Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Um Estudo de Modelos de Previsão Lineares em Séries Temporais

Guilherme N. Souza

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Orientador
Prof. Marcos Fagundes Caetano

Brasília
2020

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Um Estudo de Modelos de Previsão Lineares em Séries Temporais

Guilherme N. Souza

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Prof. Marcos Fagundes Caetano (Orientador)
CIC/UnB

Prof. Dr. Marcelo Mandelli
Universidade de Brasília

Dr. Marcus Vinicius Lamar
Universidade de Brasília

do Curso de Engenharia da Computação

Brasília, 08 de Dezembro de 2020

- -

Resumo

Modelos de previsão lineares são usados um uma variedade de cenários diferentes, alguns desses modelos dependem apenas de poucas variáveis para gerar previsões relevantes. No contexto da comunicação sem fios existe uma necessidade crescente na otimização da caracterização do usuário primário. Nesse ambiente onde a economia na transmissão de informações é importante, foi considerado o uso de modelos lineares para realizar essa caracterização. Nesse documento um estudo dos modelos AR, MA, ARMA e ARIMA é feito, de modo a explorar seus limites em três series de tempo bastante distintas, com o objetivo de compreender em quais cenários os modelos lineares poderiam ser melhor aplicados. Esse estudo é feito de modo a comparar os resultados de previsões que utilizam desses quatro métodos, evidenciando a eficácia de cada um deles.

Palavras-chave: Séries Temporais, previsão, acesso oportunístico

Abstract

Linear Prediction Models are used in a variety of different scenarios, some of these models depend only of a few variables to generate relevant predictions. On the subject of Wireless comunication there is a rising need of primary characterization optimization. Where economy in trasnmission is crucial, linear prediction models where considered to do this characterization. In this document a study on AR, MA, ARMA and ARIMA models where made, in order to explore their limits, using three datasets. This study is made by comparing the prediction results that are utilised by these four methods, pointing at the efficiency of each of them.

Keywords: Time Series, Predictions, opportunistic access

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ACF Função de Autocorrelação (do Inglês, *Autocorrelation Function*).

ANATEL Agência Nacional de Telecomunicação.

AP Ponto de Acesso (do Inglês, *Acess Point*).

AR Autoregressão (do Inglês, *Autoregressive).*

ARIMA Autoregressão com média móvel com diferenciação (do Inglês, *Autoregressive Integrated Moving Average).*

ARMA Autoregressão com média móvel (do Inglês, *Autoregressive Moving Average).*

CR Rádios Cognitivas (do Inglês, Cognitive Radio).

CSMA/CA Acesso Múltiplo com Detecção de Portadora Evitando Colisões (do Inglês, *Carrier sense multiple access with collision avoidance*).

CTS Liberação Para Receber (do Inglês, *Clear To Send*).

DIFS Espaçamento Interquadros Distribuído (do Inglês, *Distributed Inter-Frame Space*).

DSA Acesso Dinâmico ao Espectro (do Inglês, *dynamic spectrum access*).

FCC Comissão Federal de Comunicação (do Inglês *Federal Communications Commission).*

ISM Industrial, Científico e Médico(do Inglês, Industrial, Scientific and Medical).

LAN Rede Local (do Inglês, *Local Area Networks).*

MA Média Móvel (do Inglês, *Moving Average).*

MRMC Multi-rádio multi-canal (do Inglês, *Multi-Radio Multi-Channel*).

OSA Alocação Oportunística do Espectro (do Inglês, *Opportunistic Spectrum Allocation*).

PACF Função de Autocorrelação Parcial (do Inglês, *Partial Autocorrelation Function*).

PU Usuário Primário (do Inglês, *Primary User*).

RC Rádio Cognitivo (do Inglês, *Cognitive Radio*).

RTS Requerimento Para Enviar (do Inglês, *Request To Send*).

SDR Radio Definido por Software (do Inglês, *Software-defined radio).*

SIFS Espaçamento Curto Interquadros (do Inglês, *Short Inter-Frame Spacing*).

SU Usuário Secundário (do Inglês, *Secundary User*).

Capítulo 1

Introdução

O advento dos telefones celulares influenciaram na maneira em que a comunicação é feita atualmente [7]. O crescimento no número de usuários da tecnologia móvel é uma tendência [8]. Com esse contexto da comunicação sem fio em vista, se faz relevante ponderar a respeito da infraestrutura envolvida para suportar esses usuários.

A utilização do canal sem fios obedece normas que restringem e regulam o seu acesso. O controle da divisão do espectro eletromagnético é feita por órgãos reguladores. Esse ambiente sem fios se encontra, em sua maioria, dividido de maneira estática. No Brasil a Agência Nacional de Telecomunicação (ANATEL) é responsável pelo gerenciamento do espectro eletromagnético. Como é possível ver no quadro mostrado na Figura 1.1, existe uma escassez de faixas de frequência livres. A falta de faixas de frequência livres resulta em um limitador físico para novos serviços.

Um estudo feito pela Comissão Federal de Comunicação (do Inglês *Federal Communications Commission)* (FCC), o orgão que regula a divisão do espectro eletromagnético nos Estados Unidos, aponta que uma porção do espectro eletromagnético é subutilizada.

Apesar da existência de faixas no espectro que sejam não licenciadas como as de Industrial, Científico e Médico(do Inglês, Industrial, Scientific and Medical) (ISM), que são as faixas onde temos as tecnologias do WiFi e do bluetooth operando, se faz necessário uma maneira de utilizar as faixas licenciadas mais eficientemente. Em contra ponto ao modelo estático da divisão do espectro existe a tecnologia do Acesso Dinâmico ao Espectro (do Inglês, *dynamic spectrum access*) (DSA), que propõe uma divisão dinâmica. O objetivo desse tipo de acesso é aumentar a utilização do canal. A DSA possui uma variedade de modelos diferentes. Um dos modelos de acesso dinâmico é a Alocação Oportunística do Espectro (do Inglês, *Opportunistic Spectrum Allocation*) (OSA), que oferece porções do espectro não utilizadas pelo usuário licenciado, para usuários secundários.

A fim de possibilitar a comunicação, a tecnologia do rádio definido por software é empregada, com ele podemos sair do paradigma atual estático da comunicação e obter

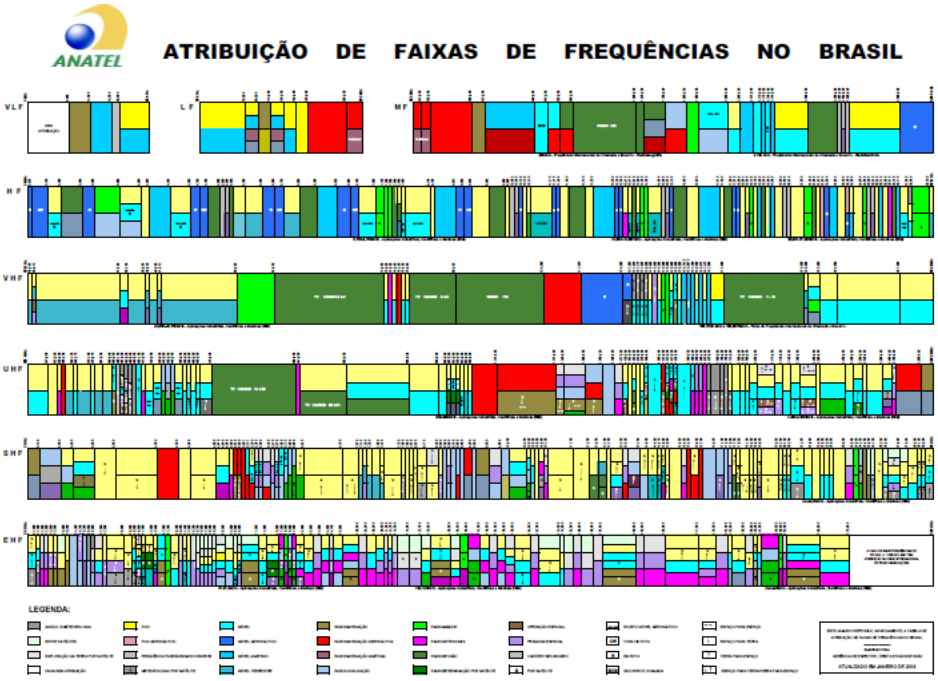

Figura 1.1: Gráfico de divisão do espectro brasileiro - ANATEL (Fonte [1]) .

um ambiente programável, mais eficiente e controlável. O uso do Rádio Cognitivo (do Inglês, *Cognitive Radio*) (RC) pode estabelecer um aumento na previsibilidade do uso do canal e com isso gerar o conforto necessário para usuário primários disponibilizarem suas bandas licenciadas para usuários não licenciados. Isso é interessante do ponto de vista financeiro, uma vez que o tempo em que o canal fica ocioso possa ser vendido, é possível proporcionar um retorno financeiro para o usuário primário, alocando esse canal durante o tempo ocioso para usuários secundários.

Dentro desse aspecto uma das preocupações é de como caracterizar o comportamento do usuário primário. Atualmente existem diversos trabalhos bem estudados na literatura que fazem esse papel, melhor explanados na Seção 2.3. Assumindo que o usuário pode possuir um comportamento periódico no uso do canal é possível pensar que modelos lineares seriam capazes de identificar esses padrões. A proposta desse texto é o estudo de modelos de previsão lineares no reconhecimento do comportamento de uso do canal do usuário primário. A aplicabilidade desse estudo é na construção de protocolos MAC que usem desse conhecimento. Permitindo assim uma melhor eficiência do canal, uma vez que possibilitaria usuários secundários a transmitirem durante a ociosidade do usuário primário.

Uma das soluções é utilizar uma análise de séries temporais para prever a existência de uma transmissão futura, baseada em um histórico de transmissões passadas. Nesse

documento utilizaremos os modelos AR, MA, ARMA e ARIMA para realizar essa previsão e antes disso vamos comparar a eficiência desses métodos em diferentes cenários, usando conjuntos de dados variados.

1.1 Problema

A escassez de faixas livres do espectro eletromagnético motiva a exploração de métodos alternativos para a divisão do mesmo. A utilização do canal de maneira mais eficiente é possível, uma vez que se é observado a ociosidade do canal. Uma das maneiras de se utilizar o canal de maneira mais eficiente é o acesso oportunístico OSA, que possuí como fase inicial a identificação do comportamento do usuário primário. Após a caracterização do usuário é possível a construção de protocolos MAC que usufruam desse conhecimento para transmitir em momentos ociosos do canal, quando o usuário primário não está transmitindo. Nesse trabalho o foco será na caracterização, e por esse motivo um estudo de viabilidade do uso de modelos lineares é feito.

1.2 Objetivo

O objetivo do trabalho é estudar o uso de Modelos Lineares para a identificação do comportamento de uso canal feita pelo usuário primário. Para tal, são modeladas três series de tempo com os modelos AR, MA, ARMA e ARIMA, onde cada uma delas possuem complexidades diferentes. Outro motivo pela escolha de modelos simples é a falta de literatura verificando a usabilidade desses métodos puros.

1.3 Metodologia

Inicialmente um levantamento de literatura foi feito, com o objetivo de identificar os modelos que seriam utilizados. A escolha das series temporais foi feita, com o foco na complexidade de previsão de cada uma delas, onde temos a primeira base de dados como sendo pouco variante e periódica, a segunda sendo mais variante e periódica e a terceira sendo mais variante e sem periodicidade definida. Foram feitas as avaliações Função de Autocorrelação (do Inglês, *Autocorrelation Function*) (ACF) e Função de Autocorrelação Parcial (do Inglês, *Partial Autocorrelation Function*) (PACF) dos conjuntos de dados de modo a auxiliar na escolha de um modelo inicial e posteriormente foram feitas alterações nesse modelo inicial com propostas de melhoras nos resultados das previsões.

1.4 Divisão do documento

Para melhor compreender a implementação do projeto e os resultados obtidos separamos o documento em 6 capítulos incluindo este introdutório. No capítulo 2 apresentamos a revisão teórica que será necessária para desenvolvermos o assunto e onde nos inspiramos para a concepção do projeto. O capítulo 3 explica a aquisição de dados, além de explanar a forma que os modelos de previsão são montados. No capítulo 4 explicamos os resultados obtidos comparando com outros estudos parecidos e no capítulo 5.1 as conclusões bem como qual foi a relevância final do projeto para a área de redes.

Capítulo 2

Revisão Teórica

Este capítulo irá descrever o conhecimento necessário para o entendimento do trabalho e toda a base teórica acumulada para a elaboração da proposta do projeto. Inicialmente nas seções a seguir são apresentadas noções básicas sobre modelos de previsão e séries temporais. Além disso o espectro eletromagnético é explicado, no sentido de justificar a escolha de uma das bases de dados usada no estudo, depois o capítulo discute mais aprofundadamente a questão da divisão dinâmica do espectro e descobriremos os pre-requisitos necessários para obtermos de maneira satisfatória a chamada oportunidade de transmissão. Outras tecnologias importantes para o entendimento do trabalho são explicadas nas Seções 2.2 e 2.1 seguintes, dentre elas o rádio cognitivo e o Wi-fi. Por fim se faz necessário mostrar outros trabalhos similares a este na Seção de estudo do estado da arte.

2.1 Análise de Séries de Tempo

Séries de tempo são definidas como um conjunto de amostras adquiridos sobre uma única variável, normalmente coletadas durante um intervalo de tempo [9]. A escolha da variável e a determinação do intervalo de amostragem são escolhidos dependendo da aplicação a ser modelada.
Exemplos de séries de tempo:

- Valores diários do preço da gasolina;
- Temperatura de um paciente medida a cada hora;
- Altura dos oceanos medida a cada ano;

Séries de tempo podem ser usadas para evidenciar padrões em diversas áreas de estudo. Muitas vezes previsões podem ser feitas utilizando os valores obtidos da série. Depois de

coletados, os valores de uma série de tempo devem ser analisados e remodelados, de modo a fazerem sentido com os dados que se quer representar.

Organizar os valores em um gráfico pode realçar propensões não notadas anteriormente e pode ajudar na escolha do método mais eficiente de previsão. As chamadas transformações alteram os valores da série de tempo transformando-as em outras com características semelhantes. Um exemplo de transformação é a Box-Cox que altera a estacionalidade da série, podendo transformar uma série não estacionaria em uma série estacionaria. Outra transformação é a diferenciação, que é capaz de diminuir ou até mesmo remover a tendência de uma série de tempo. Toda série de tempo possui três componentes pelas quais ela pode ser decomposta, cada uma delas é listada a seguir:

- Componente da sazonalidade é a que se repete com uma certo período no tempo. Exemplo seria o aumento da temperatura em uma certa época do ano a cada 12 meses ou a diminuição do fluxo de carros com a chegada do final de semana em uma certa estrada;
- Componente da tendência é o fluxo geral de uma série de tempo, para determinar a tendência se faz necessário avaliar se a média dos valores da série está subindo ou descendo;
- Componente randômica que são fatores não explícitos numa série que acabam modelando e influenciando no seu formato mas não se tem uma causalidade observável facilmente;

2.1.1 Regressão linear

Tendo em vista a importância da Regressão Linear como base para o conceito das previsões, se faz necessário elucidar como o mesmo funciona com um exemplo [10]. O exemplo montado tem o objetivo de mostrar uma regressão linear com apenas uma variável independente no contexto de séries de tempo. Levando em consideração a série de temporal apresentada em Figura 2.1 que representa o número de acessos diários a página da Wikipédia do Netflix, foram retirados 14 valores para o exemplo de regressão linear.

O objetivo da regressão é representar a série de tempo utilizando a Equação 2.1 que é a fórmula de uma reta Y_t', b é o coeficiente linear da reta, Y_t é a variável independente(nesse exemplo a própria série de tempo) e a é o coeficiente angular que determina a inclinação da reta.

$$Y_t' = b + aY_t \tag{2.1}$$

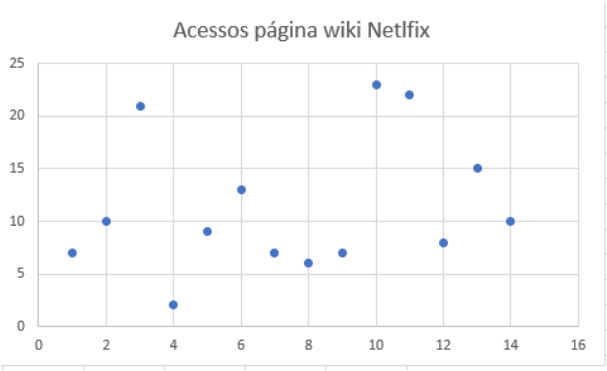

Figura 2.1: Número de acessos da Wikipédia do Netflix (Fonte [2]).

Existem uma variedade de retas que podem representar a regressão linear para um conjunto de dados, como pode ser visto na Figura 2.2, porém existe uma reta que possui um valor erro médio menor que as demais, essa será a reta escolhida nesse caso.

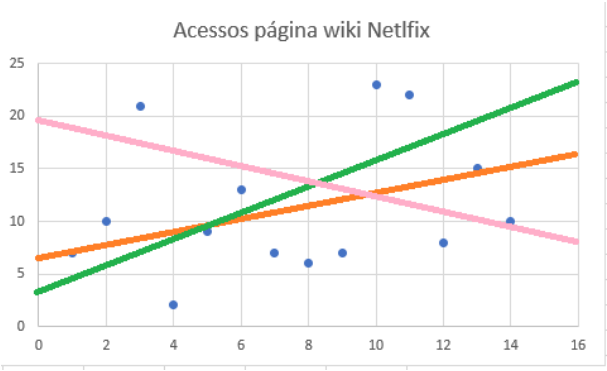

Figura 2.2: Exemplo de regressões lineares para o número de acesso da Wikipédia do Netflix.

No exemplo apresentado, a regressão linear com os coeficientes a e b que apresentam os melhores resultados são os definidos na Equação 2.2, cujo gráfico apresentado na Figura 2.3. Mesmo sendo a melhor opção de reta os valores não encaixam perfeitamente no modelo montado, e por isso existe o valor de erro para cada um dos resultados da regressão. Quando somados os valores de erro para cada valor de regressão, o resultado é exatamente o valor da série original.

$$Y_t' = 8.65 + 0.36Y_t + \varepsilon_t \tag{2.2}$$

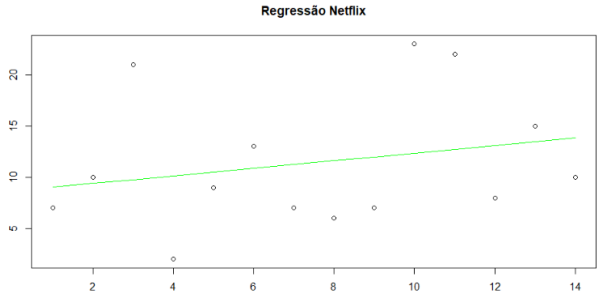

Figura 2.3: Melhor Regressão Linear exemplo Netflix Wikipage.

2.1.2 Modelo Autoregressivo

Autoregressão (do Inglês, *Autoregressive)* (AR) [10] é um modelo que assume que em uma série de tempo os valores de Y_{t+1} futuros podem ser representados pelos valores passados da série multiplicados por coeficientes. Essa relação pode ser vista com a Equação 2.3 onde ε_t representa o valor de erro e β_0 é o valor de um constante inicial e que β_1 representa um coeficiente que deve multiplicar o valor anterior da série de tempo Y_{t-1}.

Um **Lag** de uma série de tempo é a mesma série deslocada em uma posição com relação ao tempo, ou seja, o Lag_{-1} é igual a Y_{t-1}. Quando se aplica a regressão linear (explicado na Seção 2.1.1) de forma usando Y_t usando o Lag_{-1} temos uma autoregressão de primeiro nível.

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t \tag{2.3}$$

Um modelo com p igual a 3, ou seja, com valores 3 valores de *Lags* relevantes teria, uma fórmula no formato mostrado em formato Equação 2.4.

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \varepsilon_t \tag{2.4}$$

O modelo AR foi primeiro citado em [11] por Peter Whittle, juntamente com o modelo MA e que podem ser usados em conjunto.

2.1.3 Modelo Média Móvel

A Média Móvel (do Inglês, *Moving Average)* (MA) [10] é um modelo que assume que um valor futuro de uma série de tempo pode ser descrita pela soma dos valores dos erros aleatórios obtidos da regressão linear multiplicada por um coeficiente. A Equação 2.5 mostra β_0 como sendo o valor médio da série, ϕ_1 o valor do coeficiente de importância do erro e como sendo o ε_t erro.

$$Y_t = \beta_0 + \phi_1 \varepsilon_{t-1} + \varepsilon_t \tag{2.5}$$

Observando a Equação 2.6 que descreve o modelo MA. Assim como no modelo AR temos que determinar o número de fatores relevantes para cada modelo de série de tempo, no caso da média móvel assumimos o teste PACF é o suficiente para correlacionar os erros mais importantes.

$$Y_t = \beta_0 + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} ... \phi_q \varepsilon_{t-q} \tag{2.6}$$

Esse modelo linear é bastante simples quando comparado com outros modelos de previsão mas sua utilidade é notada quando os dados que se deseja prever possuem poucas informações para auxiliar na previsão.

2.1.4 ARMA

Autoregressão com média móvel (do Inglês, *Autoregressive Moving Average)* (ARMA) é a junção dos dois métodos explicados anteriormente, matematicamente a diferença entre AR e MA é a integração da série, ao modelarmos uma série assumimos que a mesma é estacionaria, muitas vezes esse não é o caso e por isso usamos a diferencial para forçar essa condição.

$$\begin{aligned} Y_t = \beta_0 + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} ... \phi_q \varepsilon_{t-q} + \\ + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} ... + \beta_q Y_{t-p} \end{aligned} \tag{2.7}$$

Avaliando a fórmula das equações apresentadas na Equação 2.7, observas-se o uso de p e q, como já explicado esses valores representam o número de Lags passados serão usados para a parte AR e o número de erros que serão usados na parte MA respectivamente. No caso da escolha de algum dos valores para p ou q sendo zero o modelo ARMA se equivale ao modelo AR quando q é zero, e MA quando o p é zero.

2.1.5 ARIMA

O modelo Autoregressão com média móvel com diferenciação (do Inglês, *Autoregressive Integrated Moving Average)* (ARIMA) apresenta uma fórmula parecida com a do modelo ARMA, a grande diferença está na aplicação de uma diferenciação na fórmula. Com o objetivo de remover a tendência de uma série de tempo algumas técnicas podem ser aplicadas para melhorar alguns casos de previsão, uma delas é a aplicada no ARIMA com a diferenciação.

É importante estabelecer que para os modelos AR, MA, ARMA e ARIMA são assumidos algumas condições, uma delas é que a série deve ser estacionária. Para garantir que a série é estacionária a média, a variância e a autocorrelação da mesma não deve variar com o passar do tempo. Encontrar essa característica em séries reais pode ser um desafio e por isso a importância do ARIMA para a previsão.

2.1.6 Autocorrelação

O conceito da autocorrelação pode ser entendido como uma relação não óbvia que existe em uma série de tempo [12]. De forma geral, é um padrão que pode ser observado entre os números da série. Explicando de maneira mais simples, comparamos cada valor da série com ele mesmo deslocado, por exemplo, o valor t comparado com os valores $t-1$, esse seria denominado o primeiro **Lag** da autocorrelação, como pode ser visto na Equação 2.8.

$$\rho(k) = \frac{Cov(X_t, X_{t+k})}{Var(X_t)} = \frac{E[(X_t - \mu)(X_{t+k} - \mu)]}{\sigma^2} \tag{2.8}$$

2.2 Protocolo IEEE 802.11g (Wi-fi)

O protocolo Wi-fi de acesso ao meio é um dos padrões de rede utilizado na LAN sem fio. Existem várias versões desse protocolo alguns deles sendo: 802.11b, 802.11a e 802.11g. Uma das diferença entre eles é a faixa de frequência em que cada um trabalha e consequentemente o modo que cada um funciona. Nosso foco será na versão 802.11g por ser a faixa mais utilizada.

Dos tipos de configurações para redes que seguem o protocolo 802.11 vamos explicar dois tipos que serão os principais: Modo de infraestrutura e modo *ad hoc* (mostrado na Figura 2.4 e Figura 2.5). A diferença básica entre eles é o tipo de suporte oferecido para a estação em que se está conectado, quando em modo de infraestrutura a estação é a responsável por estabelecer a conexão com a camada superior de redes, enquanto em modo *ad hoc* cada *host* deve fazer esse papel.

Em modo *ad hoc* a organização da arquitetura do 802.11 é diferente da normal, ao invés de termos uma centralização dos dispositivos conectados em um único AP, temos uma desorganização intrínseca nesse estilo de LAN. A comunicação descentralizada necessita de uma coordenação muito grande afim de evitar interferências nos canais e ainda continuar com uma taxa elevada de transferência.

Para explicar de maneira mais convincente a maneira pela qual é feita a comunicação, utilizaremos de um exemplo que irá reunir os conceitos mais importantes. Nesse exemplo teremos dois indivíduos conectados a uma rede 802.11, ou seja, esses indivíduos não

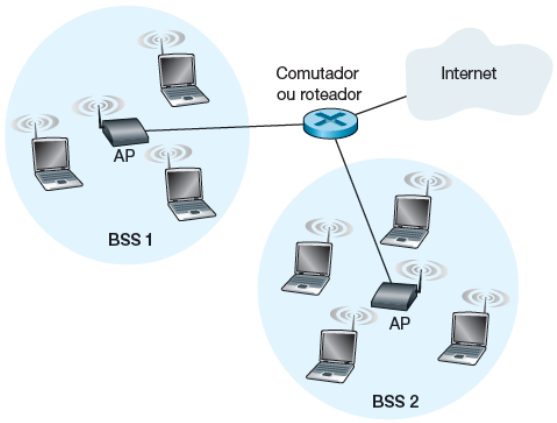

Figura 2.4: Arquitetura de LAN IEEE 802.11 (Fonte [3]).

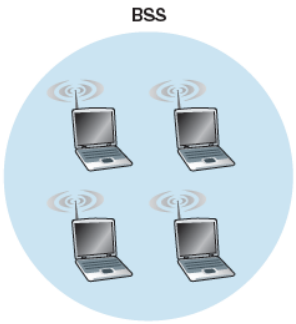

Figura 2.5: rede AD HOC IEEE 802.11(Fonte [3]).

fizeram o processo denominado de *hand-shake* e que ambos não sabem da existência do outro.

Nessa configuração de rede não existe uma maneira de reconhecer quem está transmitindo ou quando um usuário deseja transmitir. Obter esse conhecimento tende a ser muito custoso e foi necessário o desenvolvimento de técnicas para alocar o canal para a transmissão.

2.2.1 CSMA/CA

Partindo da configuração de uma rede Wi-fi formada de uma série de computadores conectados a um único ponto de acesso, podemos discutir sobre a colisão entre esses diversos computadores e o ponto de acesso. Os computadores dentro de um mesmo ponto de acesso não sabem as características de transmissão dos outros computadores da mesma rede, e mesmo assim conseguem se comunicar na mesma faixa de frequência espectral de maneira a evitar colisões. Isso é possível devido a criação de mecanismos como o CSMA/CA.

Como o Acesso Múltiplo com Detecção de Portadora Evitando Colisões (do Inglês, *Carrier sense multiple access with collision avoidance*) (CSMA/CA) é um mecanismo que faz parte do protocolo 802.11 de acesso ao meio que serve para verificar se o canal pelo qual estamos fazendo uma transmissão está disponível em redes *FullDuplex*. A maneira pela qual o protocolo funciona pode ser visto na Figura 2.6 e descrito da seguinte forma:

- Etapa 1: Inicialmente verifica-se se o canal está vazio por um determinado período de tempo chamado de Espaçamento Interquadros Distribuído (do Inglês, *Distributed Inter-Frame Space*) (DIFS).
- Etapa 2: Caso o canal se mostre ocupado, é necessário que haja mais um tempo de espera, o tempo de espera nesse caso segue a regra do recuo exponencial que será explicado mais tarde.
- Etapa 3: Depois da espera, o transmissor começa a usar o canal enviando um pacote, e logo em seguida espera uma mensagem de *acknowledge* do ponto de acesso. Esse tempo de espera pelo reconhecimento é chamado de Espaçamento Curto Interquadros (do Inglês, *Short Inter-Frame Spacing*) (SIFS)
- Etapa 4: Caso não tenha acontecido uma colisão a transmissão foi um sucesso e o protocolo se repete a partir da etapa 2.

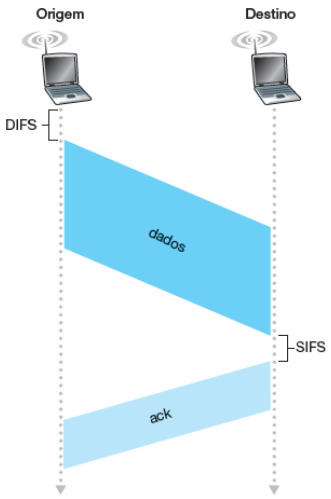

Figura 2.6: Funcionamento do CSMA/CA (Fonte [3]).

Esse processo se repete até que toda a transmissão seja satisfatória, partindo do ponto em que usamos o recuo exponencial. O tempo escolhido nessa etapa é delimitado por $2^n - 1$, onde temos um número escolhido de maneira aleatória que vai de 0 até n, com n sendo o número de pacotes a serem transmitidos.

A aleatoriedade na escolha de n segue a lógica de que no caso de duas estações almejarem utilizar um canal ocioso ao mesmo tempo, ainda teriam que esperar um *backoff* aleatório que provavelmente não seria o mesmo da estação rival. Nesse exemplo teríamos uma estação que transmitiria mais cedo utilizando do canal e o ocupando, fazendo com que a estação perdedora congele a contagem do seu próprio *backoff* e esperando que a transmissão da vencedora termine.

Outro motivo pelo qual mudamos o método de tratar colisões é devido ao meio pelo qual as informações trafegam. No CSMA/CA a via de transmissão são ondas Wi-fi, por isso o devemos tentar evitar colisões ao invés de tratar no caso de uma ocorrência delas, se fossem fios o tratamento seria diferente.

Outra possibilidade que também é utilizada para enviar pacotes em uma rede Wi-fi é a chamada de RTS e CTS. Esse método é usado de maneira opcional para evitar colisões em redes sem fio.Quando temos um pacote de dados muito grande a ser transmitido, se faz necessário alocar o canal. Pacotes menores podem correr o risco de se perderem, em contra partida, pacotes grades consomem muitos recursos para valer a pena correr esse risco. É importante salientar que o objetivo principal do protocolo é de verificar se o canal se encontra livre para ambos em uma janela de tempo específica de modo assegurar que nenhuma outra transmissão esteja sendo feita naquele momento, no contexto do acesso oportunístico essa verificação estaria sendo feita inicialmente por um Usuário Secundário (do Inglês, *Secundary User*) (SU) transmissor(verifica a existência de Usuário Primário (do Inglês, *Primary User*) (PU)s na sua área) que após a verificação envia uma mensagem do tipo Requerimento Para Enviar (do Inglês, *Request To Send*) (RTS) para um SU recebedor, este ao receber a mensagem sabe que o canal está livre para transmissão. Depois o SU recebedor envia uma mensagem do tipo Liberação Para Receber (do Inglês, *Clear To Send*) (CTS) para seu transmissor com o objetivo de avisar sobre a disponibilidade do canal. No recebimento do CTS uma conexão é formada e a troca de dados é possibilitada para os dois SUs em questão.

A Figura 2.7 mostra um exemplo de comunicação utilizando RTS-CTS evidenciando o que acontece com os outros nós que não participam da comunicação principal, como o CTS é ouvido por todos os nós de um mesmo Ponto de Acesso (do Inglês, *Acess Point*) (AP) o recebimento de um CTS não originado por você significa que algum outro nó recebeu a permissão de envio de dados, por isso devemos esperar até que o envio termine, ou seja, até que recebamos um *ack* do AP.

A troca de RTS-CTS é muito boa para estabelecer a diminuição de colisões entre usuários secundários também, quando um outro usuário secundário escuta a transmissão de mensagens RTS-CTS que não são destinadas para ele, o mesmo já sabe que existem outros dois nós comunicando entre si. O que ocorre é que o canal deixa de estar vago e

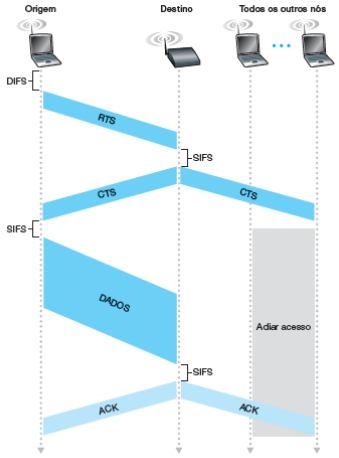

Figura 2.7: Exemplo de conexão via RTS-CTS (Fonte [3]).

isso impossibilita a transmissão para novos usuários secundários naquela janela de tempo.

2.3 Acesso oportunístico do espectro

Um dos métodos utilizados para resolver o problema de alocação do espectro é o Alocação Oportunística do Espectro (do Inglês, *Opportunistic Spectrum Allocation*), com o objetivo de preencher as faixas disponíveis nas transmissões dos PUs efetuando a detecção dos padrões de transmissão, segundo escolhendo o que se deve fazer para evitar interferências e por último tomar as medidas para utilização da parte ociosa do canal.

Para compreender sobre o que se consiste a OSA, podemos dividir o processo em três partes bem diferenciadas: identificação de oportunidade, exploração da oportunidade e política regulamentadora [13]. A identificação de oportunidade pode ser definida como os métodos pelos quais identificamos as frequências não utilizadas no espectro, exemplo disso pode ser um usuário secundário prevendo padrões na utilização do canal por um usuário primário. Para isso devemos verificar se o canal está livre por meio do sensoriamento espectral que detalharemos futuramente na Seção 2.3.2.

Já a exploração das oportunidades é a maneira pela qual as informações dadas pela identificação explicada na etapa anterior e efetivamente ocupar a faixa vazia do espectro não utilizada. Essa exploração deve ser feita de modo a não comprometer as restrições estabelecidas pelo política regulamentadora que tenta proteger os usuários primários.

Dentro das dificuldades observadas na área do acesso dinâmico ao meio, temos a escolha das oportunidades de acesso como sendo uma das maiores, o problema reside em decidir com eficiência quando o usuário secundário pode transmitir, em qual canal e por quanto tempo transmitir sem atrapalhar o acesso do usuário primário. Alguns outros

pontos também são requisitados para que tenhamos um protocolo OSA MAC, que atenda as necessidades dos usuário secundários atuais podemos citar

- As decisões devem ser tomadas em tempo real para determinar em qual parte do espectro devemos medir e atuar;
- Os nós devem determinar qual o melhor canal para se transmitir baseados nas saídas do protocolo;
- Os nós devem ter uma coordenação com relação a utilização do canal;
- Os nós devem deixar de utilizar o canal na presença de um usuário licenciado.

2.3.1 Detecção de Oportunidades

Inicialmente se faz necessário explicitar um modelo simples de comunicação contendo um PU e um SU como mostrado na Figura 2.8.

Quando estamos na fase de detecção de oportunidades devemos ter em mente os pré-requisitos que efetivamente vão categorizar aquele espaço como uma chance de transmissão efetivamente dita. Alguns deles sendo a interferência permitida a um transmissor SU efetuar num receptor PU. Outra decisão é, o que deve ser feito no caso de dois SUs estiverem com a necessidade de transmissão e detectarem uma oportunidade. Existem uma série de fatores que precisam ser preenchidos para essa caracterização e é por isso que temos um protocolo a seguir para detectar oportunidades e o que fazer depois de encontrada. No exemplo, temos dois SUs (A e B), que desejam se comunicar. SU 'A' deseja se comunicar com 'B' utilizando o canal não licenciado, nenhum dos dois sabe os padrões de comunicação dos PUs da região. 'A' só poderá fazer isso caso sua transmissão não interfira com nenhum PU que esteja tentando transmitir e caso 'B' não interfira no recebimento de nenhum usuário primário tentando receber a transmissão.

Na Figura 2.8, r_{tx} representa o raio máximo em que não se deve existir PUs receptores naquela região, desse modo caracterizando uma oportunidade de transmissão pelo ponto de vista do transmissor 'A'. De modo similar devemos obter uma oportunidade pelo lado do receptor 'B', onde o raio r_{rx} representa a distância que respeitaremos a existência de transmissores primários. Caso ambos requisitos sejam atendidos existe uma oportunidade de transmissão. Os valores de r_{tx} e r_{rx} variam de acordo com a potência dos transmissores de PU e de SU respectivamente [4].

Tendo estabelecidas os pré-requisitos passamos então pela forma que efetuamos a detecção propriamente dita. As maneiras pelas quais podemos detectar as oportunidades pode são diversas. Usando o modelo da figura podemos demonstrar um dos protocolos mais simples.

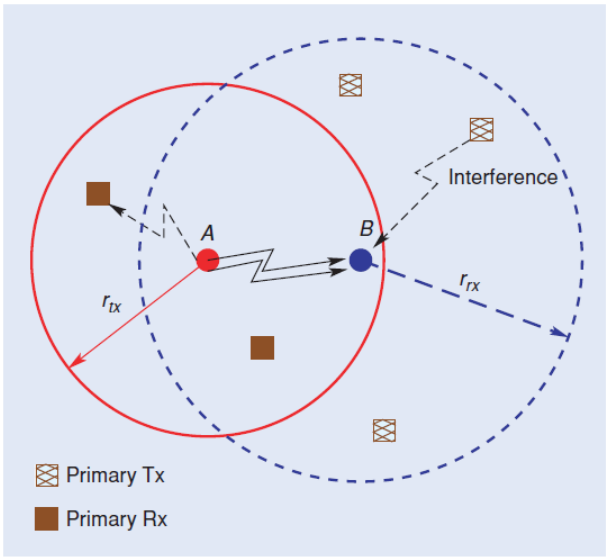

Figura 2.8: Exemplo de transmissão e possibilidade de oportunidade (Fonte [4]).

2.3.2 Sensoriamento Espectral

Afim de podermos reutilizar as faixas subutilizadas do espectro eletromagnético, precisamos inicialmente saber se aquela determinada faixa se encontra vazia. Essa prática de detecção do espectro é realizada pelo rádio cognitivo com o objetivo de obter informações a respeito do canal com relação a seu estado atual e pode ser feita das seguintes maneiras por exemplo: Detecção de Energia, o Formato de Onda, o Ciclo estacionário, a Identificação de Rádio e a Filtragem [14].

O método mais simples de sensoriamento espectral é o de detecção de energia, uma vez que a sua implementação requer pouco poder computacional devido a simplicidade comparado aos outro métodos. Basicamente quando se quer saber sobre a situação atual de uma faixa do espectro, é possível verificar o espectro de frequência e medir a sua utilização por meio da detecção dos níveis de energia. Com essa medição podemos então comparar os valores adquiridos com um valor limite previamente estabelecido, qualquer medição que esteja abaixo do limite é considerado ruído e as medições acima do limite são consideradas transmissões, ou seja, a faixa de frequência está em utilização por um outro transmissor.

Estipular o valor para esse limite a princípio parece uma tarefa amena, porém não é isso que a realidade mostra. No caso de valores muito elevados para o limite teremos como resultado uma maior chance de caracterizar uma transmissão de PU como sendo um ruído, enquanto um limite baixo teremos ruídos que serão caracterizados como transmissões abaixando a eficiência de reutilização do espectro.

Note que existe uma vantagem nessa técnica com relação a quantidade de informação

necessária para sensoriar o canal, visto que não é necessário conhecimento prévio do modo de operação do PU. Um problema na detecção por energia está na inabilidade do algoritmo em perceber a diferença entre a transmissão de um PU com a detecção de um ruído, a melhora na eficiência nesse algoritmo muitas vezes se baseia na escolha entre o que vai ser considerado ruído e o que será considerado transmissão no PU[15], sem falar de problemas como o terminal oculto e em casos de canais com transmissões não ideais.

O sensoriamento espectral é a tarefa mais importante realizada pelo rádio cognitivo, e deve ser feita de maneira eficiente para melhorar a taxa de utilização do espectro. Sabendo o estado atual do espectro devemos então decidir se durante aquele período de tempo o SU pode ou não utilizar o canal, se a brecha temporal caracteriza uma oportunidade de transmissão.

Dado o conceito básico de uma oportunidade o próximo passo é descobrir qual o potencial máximo de interferência permitido pelas políticas de interferência. A maneira pela qual escolhemos abordar a divisão do espectro e sua política regulatória é sem dúvida um desafio. Um dos pontos da discussão é o quão rígidos devemos ser ao abordar a questão do cedimento de espaço em um canal, e quais os parâmetros mínimos uma política regulatória deve possuir de modo a permitir boas oportunidades para SUs e garantir a segurança de PUs.

2.3.3 Interferência

Interência é um fenômeno físico que acontece ao se observar ondas de frequências parecidas trafegando em um mesmo ambiente ao mesmo tempo. No contexto da comunicação esse fenômeno é prejudicial. Evitar interferência no canal de transmissão é de muita importância no processo de reutilização do espectro. No modelo que estamos assumindo o PU não deve ser prejudicado em momento nenhum de sua transmissão, uma vez que ele é o detentor dos privilégios daquela canal. A grande questão reside no que vamos definir como "prejudicado", muitas vezes o órgão regulador é quem define a taxa pela qual é aceitável um SU interfira na transmissão do PU. O conjunto de definições desse tipo define uma política de interferência. Nesse caso o mínimo necessário para obtermos uma política de interferência satisfatória são dois parâmetros: o potencial máximo de interferência admitido em um PU recebedor e o número total de colisões admitidas em uma janela de tempo.

2.4 Rádio definido por Software

A necessidade por um modo de controlar de maneira mais eficiente o ambiente altamente dinâmico das redes de comunicação móvel, culminou na crescente demanda por dispositi-

vos como os rádios definidos por software. A maneira pela qual obtemos o sensoriamento do espectro eletromagnético foi facilitada pela utilização de tais dispositivos.

2.4.1 Rádio Cognitivo

Uma definição de Rádio Cognitivo (CR) foi feita por Arslan Hüseyin [16], e pode ser entendida como uma rede capaz de observar o ambiente em que está inserida e através dessas observações tomar decisões que melhorem sua performance. O uso da palavra cognitivo traz justamente a ideia de que é um rádio com a capacidade de percepção aprimorada, ele usa os conhecimentos adquiridos pelo sensoriamento e pelo software para a tomada de decisões. Essa capacidade é de importância fundamental para o acesso dinâmico ao meio, além disso os rádios cognitivos possuem outros usos em outras áreas do conhecimento.

A melhora na performance deve compensar os custos adquiridos pela adoção da tecnologia, uma vez que existe um aumento na complexidade quando comparamos rádios convencionais com rádios cognitivos. Os aumentos procedem dos custos relacionados para manter o rádio funcionando e ao custo de processamento adicional chamado de *overhead*. Além disso rádios cognitivos devem melhorar a taxa de transmissão e segurança, com a intenção de prever o comportamento do canal ao invés de simplesmente reagir as mudanças nas entradas da rede.

Arquitetura dos CR

O funcionamento de um CR é semelhante ao de um rádio quando comparamos a captação e transmissão para múltiplos canais. A maior diferença reside em sua capacidade de reconfiguração por software, isso é possível devido a um certo número de medidas que o rádio cognitivo é capaz de fazer, dando a oportunidade para a escolha de qual o melhor protocolo para uma situação específica. Além disso o CR pode ser usado para um leque de situação além do preenchimento do espectro, algumas delas sendo o monitoramento, segurança e outras áreas da comunicação.

Beacons

Os beacons são aparelhos que servem para transmitir alguma informação de maneira constante. Normalmente essa tecnologia é usada para informar de maneira periódica a localização de um determinado a dispositivo.

2.5 Revisão do Estado da Arte

Estudos feitos na área de alocação do espectro Wi-fi diversos são os resultados adquiridos, uma vez que existem um grande número de pesquisadores tentando melhorar esse problema. O atual impasse pelo qual o futuro do espectro se encontra é representado pelo aumento na quantidade de dispositivos móveis e pela busca de aplicações que necessitam da utilização do espectro para se manterem funcionais. Por outro lado o modo pelo qual fazemos a distribuição das faixas desse espectro, de maneira estática e por muitas vezes ineficiente.

Existem duas maneiras principais que para a resolução do problema do espectro. A primeira delas é o Multi-rádio multi-canal (do Inglês, *multi-radio multi-channel*) (MRMC), que consiste da atribuição de vários canais de rádio a cada um dos nós de transmissão, de modo que existe uma sobreposição na transmissão. O outro método é o que usaremos nesse artigo, consiste do uso do espectro definido por um software, capaz de decidir pelo DSA as melhores maneiras para utilizar o canal.

Um ponto importante quando se trata do modelo de previsão é a complexidade do que se tenta prever. Muitas vezes o modelo ARIMA e outros métodos lineares são usados para dados mais simples e que existem poucas informações que possam ser relacionadas para a melhora da previsão. Artigos que obtém sucesso no uso do ARIMA usam séries simples ou em conjunto com um outro modelo de previsão mais complexo. Houve uma dificuldade para achar artigos com o estudo de modelos lineares aplicados sem auxilio de outros métodos. Isso motiva o estudo feito nesse documento, uma vez que são exploradas várias previsões que usam somente o AR, MA, ARMA e ARIMA.

2.5.1 Previsão do comportamento de usuário aplicado a Smart Homes

Motivado pelo objetivo de melhorar a experiência de um usuário de casas *smart* [17], o artigo proposto tenta prever o horário de acionamento do aquecedor de água todos os dias durante 12 semanas e tenta prever a décima terceira semana. Com um total de 84 amostras um modelo ARIMA aprimorado é usado, e comparado com resultados de um modelo ARIMA normal. Resultados indicam que o ARIMA melhorado obtém melhores resultados para previsões de curto prazo.

2.5.2 Modelagem de Usuário primário usando um modelo Hibrido ARIMA/NARX

Explorando o assunto acesso dinâmico da rede pelo usuário secundário usando rádios cognitivos o trabalho tenta prever o comportamento de acesso do usuário primário usando um modelo híbrido [18].

Esse modelo de previsão híbrido é composto de um modelo ARIMA cujos valores de p e q são usados para determinar a complexidade do modelo NARX. Nesse estudo resultados de MSE do modelo híbrido se mostraram 23,23% menor que o modelo ARIMA puro e 29,25% menor que o modelo usando Rede Neurais aleatórios.

2.5.3 Previsão em Séries de Tempo usando um modelo híbrido ARIMA e Rede Neural

Quando se une um modelo ARIMA com características lineares com um modelo Neural com características não lineares a previsão para modelos reais complexos tendem a ter ambas, essa é a motivação desse artigo [19].

O estudo mostra que existe um modelo mais geral que combina duas técnicas diferentes para uso com dados reais. A dificuldade em determinar o melhor modelo para cada caso, desmotiva o uso de técnicas de regressão linear. O artigo apresenta resultados do uso dos dois modelos, linear e não-linear, em conjunto.

Capítulo 3

Metodologia

A organização do capítulo será feita da seguinte forma. Serão mostrados estudos feitos em duas bases de dados diferentes, uma sendo uma série de tempo que conta o número de passageiros aéreos a cada mês e outra sendo uma série de tempo contendo o número de acessos à página da Wikipédia da Netflix. Em seguida será mostrada como foi feita aquisição do conjunto de dados referentes a transmissão Wi-fi, são mostrados os estudos dos modelos nesses conjuntos de dados. Em cada uma das etapas é feito um teste PACF e ACF, um passo comum na análise de séries de tempos. Outras características das séries estudadas também são expostas, como a estacionariedade da série, se existe uma tendência óbvia ou não, se existe uma sazonalidade óbvia ou não.

Com o objetivo de testar a eficiência dos modelos matemáticos para previsão de séries de tempo foram efetuados previsões para 4 modelos diferentes AR, MA, ARMA e ARIMA. Todas os resultados foram feitos utilizando a linguagem R e na IDE RStudio. Outras previsões foram feitas usando modelos prontos da biblioteca *forecast* do RStudio.

3.1 Comparando modelos usando MSE, MAD e MPE

Em outra parte do estudo é realizada a comparação de modelos com valores p e q diferentes e seus resultados [20]. A ideia por trás dessa comparação é avaliar qual seria o melhor modelo de previsão possível para o conjunto de dados estudado, e saber os limites dos preditores lineares. Uma vez entendidas as dificuldades atreladas a prever séries temporais com propriedades diversas, se faz possível imaginar aplicabilidades para esse conhecimento adquirido. Existe uma variedade de medidas possíveis para melhorar os resultados das previsões de um modelo linear, mas também existe um limiar que delimita se vale a pena usar esses métodos. Por esses motivos alterações com relação à base de dados são feitas. Algumas dessas transformações mais complexas são citadas na Seção 2.1, mas outras

alterações mais simples também são feitas, referente a maneira que a série temporal é disposta.

Nessa etapa são produzidas tabelas com os resultados das previsões, contendo o erro quadrático médio(MSE), desvio absoluto médio(MAD) e erro percentual médio(MPE).

MSE

O erro médio quadrático é uma métrica comumente usada na área de análise de dados, e também é bastante comum para medir a precisão de modelos lineares. Quando avaliamos os dados com essa métrica é importante levar em consideração o conjunto de dados que está sendo previsto, uma vez que depende fundamentalmente da variância dos dados previstos. Outra vantagem dessa métrica é excluir erros negativos, uma vez que previsões são elevadas ao quadrado e depois tirada a média.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i' - Y_i)^2 \tag{3.1}$$

MAD

O MAD é capaz de informar a distância de cada um dos pontos de um conjunto de dados e a média. Isso serve para informar o quanto esse conjunto de dados é uniforme.

$$MAD = \frac{1}{n}\sum_{j=1}^{n}|y_j - y_j'| \tag{3.2}$$

MPE

O MPE é o erro médio percentual, e diferente das outra métricas, ela leva em conta a grandeza do dado avaliado. Essa métrica é dada em porcentagem e pode ser usada para comparar diferentes conjuntos de dados uma vez que seu resultado não depende do tamanho dos dados.

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i' - Y_i)}{Y_i} \times 100 \tag{3.3}$$

3.2 Previsões do número de Passageiros Aéreos

Como base para os testes foram utilizados os arquivos padrões do RStudio, o primeiro arquivo analisado é o intitulado *Airpassengers* [5] que mostra os números referentes ao total mensal de passageiros de companhias aéreas internacionais entre 1949 a 1960, Figura 3.1.

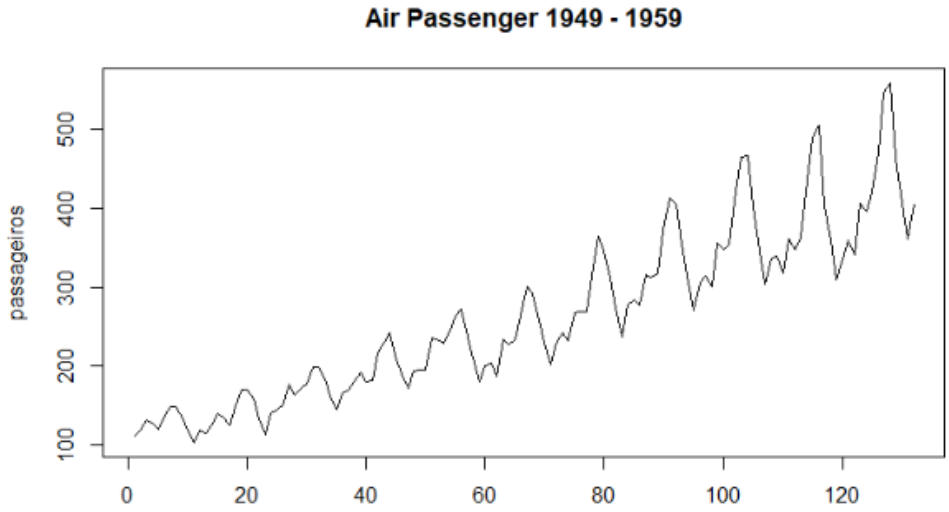

Figura 3.1: Número de Passageiros 1949 até 1959 [5].

Esse conjunto de dados é clássico em estudos que envolvem séries temporais, esse é um dos motivos pelo qual ele foi escolhido. Outro motivo é a facilidade pela qual alguns elementos importantes para previsões lineares podem ser notados. Exemplo disso é tendência de crescida da série, e a sazonalidade anual.

3.2.1 Previsões do número de Acessos

Esse conjunto de dados representa o número de acessos diários as páginas do Wikipédia, das diversas páginas disponíveis foi escolhida a do Netflix, mostrada na Figura 3.2 [21]. Neste trabalho, realizamos a comparação do modelo AR(10) com outros métodos lineares, a fim de validar os resultados descritos em [6].

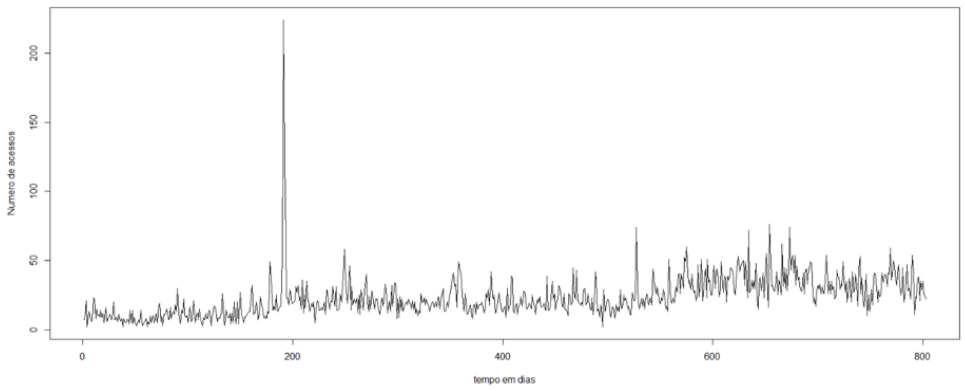

Figura 3.2: Número de Acessos à página Wikipedia Netflix.

Outro motivo importante por trás da escolha desse conjunto de dados é a complexidade. Apesar da forte presença de uma sazonalidade semanal, a série é mais variável que a *Airpassengers*. Além disso a série possui picos de acessos em alguns momentos, que dificulta as previsões lineares no que diz respeito a identificação de correlações.

3.2.2 Previsões dos Slots Idle

O ultimo conjunto de dados é o mais complexo. Usando os valores adquiridos na captura de potência de banda feitas na área da Unb por um rádio cognitivo, foi possível montar uma série temporal com 131.071 valores de potência em Watts.

Esses dados são agrupados de diferentes formas, montando séries temporais diferentes, tentando sempre construir uma série que possa ser mais facilmente prevista. As alterações feitas são explicadas com o passar do capítulo mas a série final possui um total de 83 valores e é apresentada na Figura 3.3. Assim como nos outros conjuntos de dados algumas análises são feitas para tentar caracterizar a série. A seguir uma variedade de modelos lineares é montada e comparada, tentando obter uma previsão com valores de erros cada vez menores.

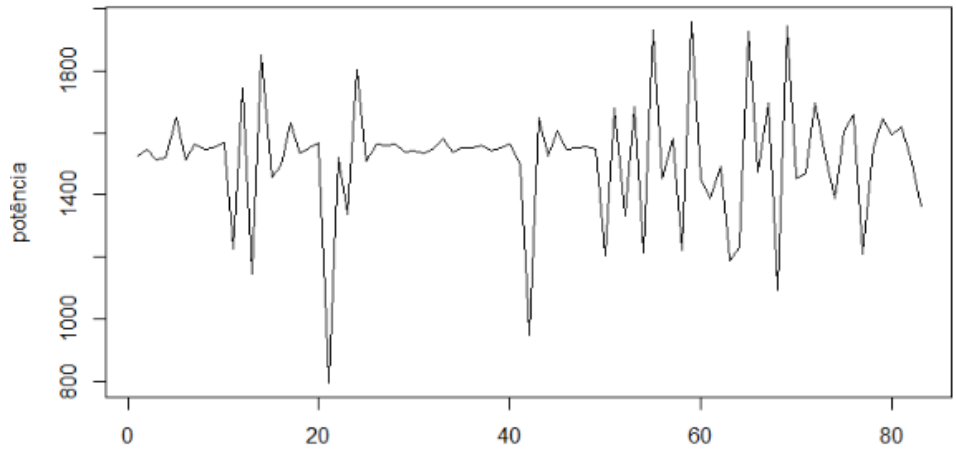

Figura 3.3: Amostra de energia sem Beacons.

3.3 Aquisição e Processamento de Sinais

Neste trabalho, a captura das transmissões foi feita utilizando Radio Definido por Software (do Inglês, *Software-defined radio)* (SDR), com o objetivo de obtermos os valores de energia do sinal. Foi necessário inicialmente adquirir os valores de fase (I) e de qua-

dratura(Q). Com esses dois elementos e utilizando a Equação 3.4 é possível calcular os valores de energia de janelas de tempo (N) amostras.

$$E_n = \frac{1}{N} \sum_{j=t}^{t+N} I(j)^2 + Q(j)^2, \tag{3.4}$$

Utilizou-se um USRP desenvolvido pela Ettus Research do modelo N210, com capacidade de amostragem de 400MHz com resolução de 16 bits. A versão modular do USRP é usada para configurar uma rede IEEE no modo de ativação *ad-hoc* 802.11g utilizando o CSMA/CA como base de transmissão.

Como especificado pelo padrão 802.11 a largura de banda é de 22MHz, a taxa de amostragem foi de $f = 25$MHz, portanto T com o período de 40ns. Esse período é adequado para englobar detectar o tempo de SIFS, portanto, é adequado. A frequência central foi de 2,47225GHz correspondente ao canal 13 do padrão IEEE 802.11. Os resultados da amostragem são armazenados em dois sinais ortogonais $I(t)$ e $Q(t)$ seguindo o padrão IEEE 754 de ponto flutuante preciso simples em formato binário.

O valor de N, referente ao número de amostras, foi escolhido como sendo 128 amostras. O objetivo para essa escolha é devido ao menor tempo que deve ser detectado, o tempo de SIFS, que é de 9us. Assim $N = 25MHz \times 9\mu s = 225$, sendo a maior potência de 2 menor que 225 igual a 128 amostras.

3.3.1 Limiarização

O sinal obtido é limiarizado, de modo a caracterizar uma transmissão ou um tempo de espera, por esse motivo só foi preciso escolher um valor pelos quais todos os outros são caracterizados. O valor de limiar foi escolhido como sendo igual a 10^{-7}, mostrado na Figura 3.4 pela linha pontilhada preta. O sinal limiarizado é apresentado em azul nesta figura.

Como o valor de *alpha* varia dependendo do ambiente no qual se fazem as capturas, foi necessário escolher um método para determinar o valor de *alpha*, no caso o *k-means* é o método de *clusterização* que utiliza um número K de centroides para reduzir o número de perdas do sinal. Observando a Figura 3.5 podemos observar um exemplo da utilização do *k-means*.

O agrupamento mostrado foi eficiente para remover valores que saiam muito da média do grupo, o que facilitou na avaliação de uma transmissão, uma vez que o valor obtido pode ter tido um pequeno erro ao ser calculado.

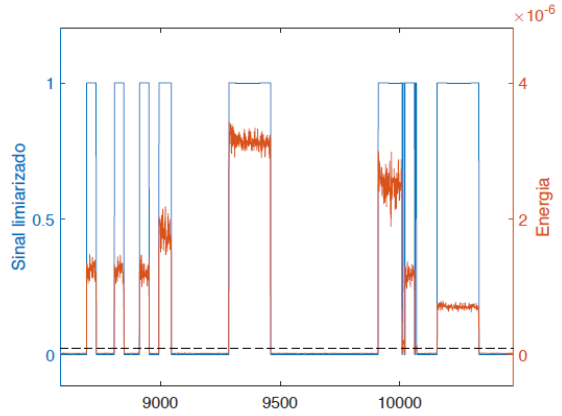

Figura 3.4: Limiarização da função de energia..

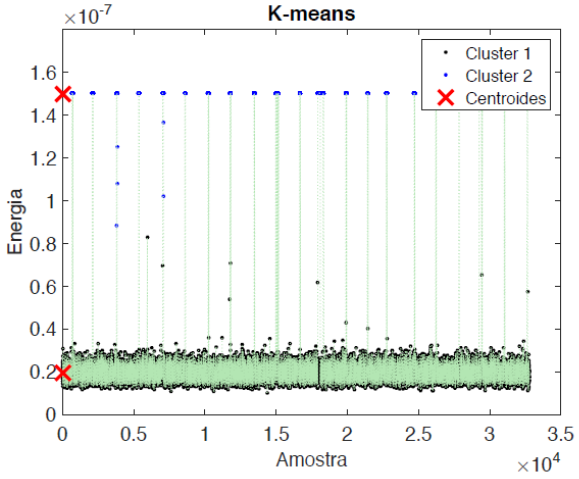

Figura 3.5: Exemplo de utilização de *k-means*.

Capítulo 4

Resultados Experimentais

Esse capítulo irá mostrar os resultados referentes aos três conjuntos de dados. Os conhecimentos adquiridos na exploração dos conjuntos mais simples são aplicados nos conjuntos de dados mais complexos. Lembrando que é interessante tentar alcançar o melhor resultado possível para cada cenário, uma vez que os limites dos preditores lineares podem ser explorados.

4.1 Resultados Iniciais *Airpassangers*

Para validarmos os modelos de previsão retiramos da amostra os valores de 1960 e tentamos fazer a previsão para aquele ano, inicialmente obteve-se os testes ACF e PACF na série de tempo e os resultados são mostrados na Figura 4.1 e na Figura 4.2. Efetuar os testes não é uma medida obrigatória, na realidade é apenas um método comumente utilizado em previsões.

Os testes servem para procurar alguma correlação entre os valores da série de tempo, e talvez evidenciar uma tendência que não pode ser avaliada pelos valores puros da mesma. É importante salientar que a periodicidade escolhida para esses teste foi de 12 amostras, isso se justifica devida a própria característica da série do tempo. Os dados que compõe a série foram colhidos num espaço de 1 mês, totalizando 12 para cada ano, razoável pensarmos que a sazonalidade se repita com essa mesma frequência. Apoiando a previsão nessa componente sazonal ajuda no resultado final da previsão.

Outro fator de muita importância é a transformação feita nos valores da série de tempo. Assim como foi explicado na Seção 2.1.2, muitas vezes é necessário efetuar algumas modificações nos valores da série para que uma boa previsão seja feita. No caso da série *Airpassengers* os valores puros quando avaliados pelos testes ACF e PACF refletem os resultados mostrados nas Figuras 4.1 e 4.2.

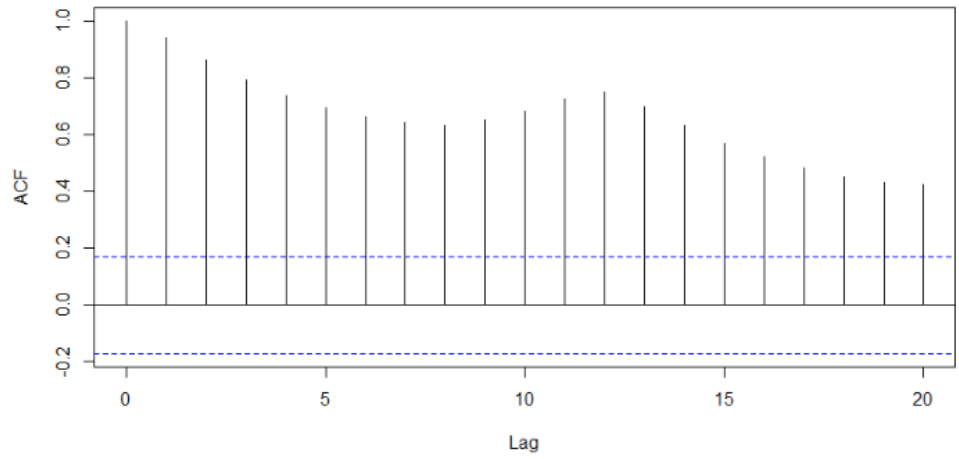

Figura 4.1: Teste ACF passageiros aéreos.

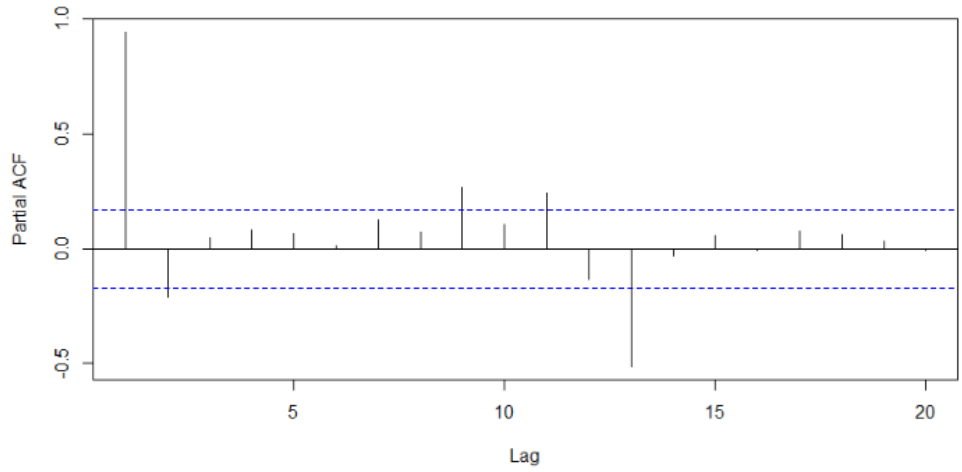

Figura 4.2: Teste PACF passageiros aéreos.

Nesse caso não é possível determinar a correlação entre os diferentes números da série uma vez que o teste mostra valores muito altos de correlação para todos os valores futuros de n no teste ACF da Figura 4.1. Para efetuarmos uma boa previsão na série ela precisa ter uma distribuição do tipo estacionária, ou seja, quando temos um valor médio somados com um valor aleatório ao redor dessa variável. Valores obtidos de ambientes reais normalmente contém partes estacionárias e partes não estacionárias, e quando passamos os valores por uma diferencial certificamos que a série é na sua totalidade estacionária.

Quando observamos o teste ACF mostrado na Figura 4.3 e na Figura 4.4 feito com

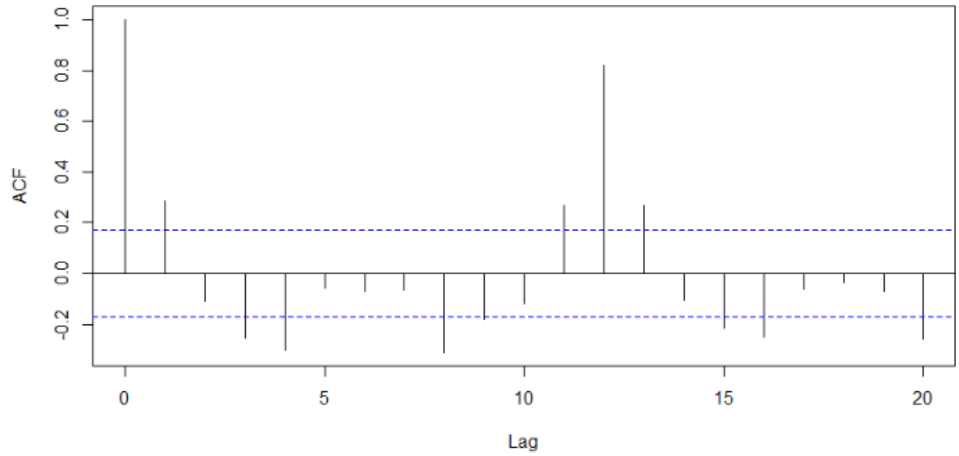

Figura 4.3: Teste ACF passageiros aéreos com diferenciação.

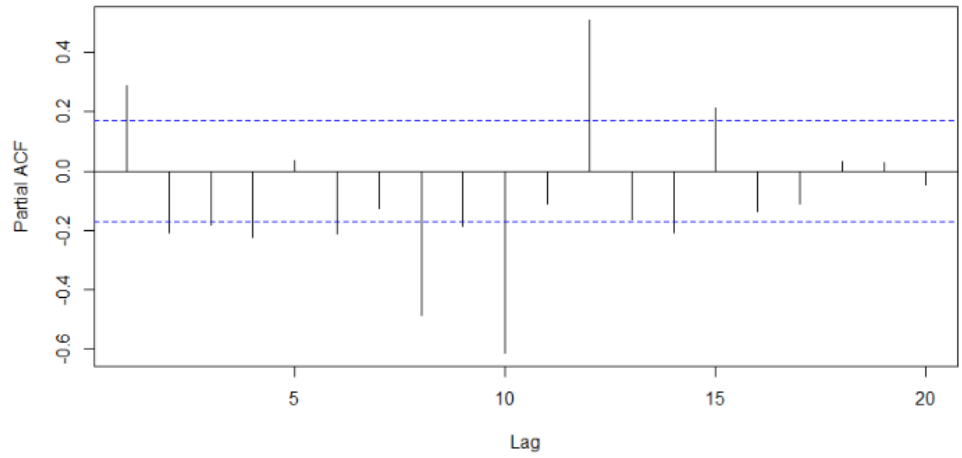

Figura 4.4: Teste PACF passageiros aéreos com diferenciação.

diferenciação, podemos avaliar mais facilmente a série de tempo. Quando observamos os valores de maior relevância no teste de correlação podemos ver que o valor de $n1$ é igual a 1 de correlação, isso se dá pois um termo sempre é altamente correlacionado com ele mesmo. O que realmente queremos observar são os próximos números, uma vez que valores altos de correlação provavelmente vão impactar na previsão ARIMA. Os valores acima de 0.2 de correlação para o teste ACF são escolhidos para o valor do índice q, referente ao valor de MA, nesse caso iguais a 1. Enquanto aos valores obtidos do teste PACF são responsáveis para guiar na escolha dos valores de p, na parte AR do ARIMA,

nesse caso sendo 0.

Podemos visualizar valores de alta correlação em lags futuros como de $n11$, $n12$ e $n13$ isso acontece exatamente pela frequência explicada anteriormente, a cada 12 valores de amostra teremos essa mesma tendência se repetindo a cada ano.

Utilizando os valores coletados da análise dos testes obtemos um modelo ARIMA(0,1,1), observe a Figura 4.5, que na realidade é um modelo MA com a utilização da diferencial, uma vez que a parte AR está com a os valores zerados de p. O gráfico mostra em preto os valores originais da amostra de *Airpassengers* e em verde os valores previstos pelo modelo.

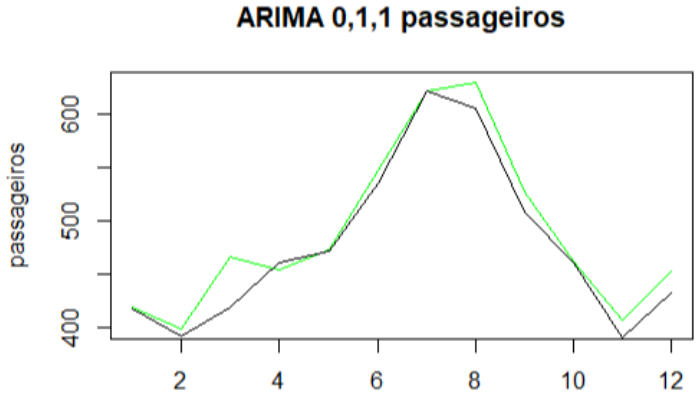

Figura 4.5: Previsão de passageiros aéreos 1961 - ARIMA 011.

Tabela 4.1: Resultados ARIMA 0,1,1 com frequência 12

	MSE	MAD	MPE
ARIMA 0,1,1	444,80	16,21	3,65

Tabela 4.2: Resultados AR AirPassengers

	MSE	MAS	MPE
AR1	291,44	12,65	2,75
AR2	290,78	12,5	2,76
AR3	292,59	12,49	2,76
AR5	315,85	13,46	3,01
AR10	327,03	13,82	3,11
AR20	274,8	11,85	2,58
AR30	272,08	12,89	2,71
AR40	136,4	9,37	1,98
AR50	124,81	8,82	1,89
AR53	114,47	8,78	1,9

Comparando modelos Airpassengers

Muitas vezes os testes feitos tentando apresentar o melhor modelo não leva em consideração uma variedade de características que no fim não apresentam o melhor modelo real.

De maneira mais empírica foram feitos testes com p, d e q com diferentes valores e apresentados com seus valores de MSE, MAD E MPE. O objetivo é poder averiguar os limites do modelo ARIMA para os diferentes dados.

A Tabela 4.2 mostra os modelos AR que apresentam uma melhora nos resultados tendo em vista a diminuição dos valores MSE, essa melhora segue com o aumento do números de lags p usados no modelo. Uma interpretação possível para essa melhora é que os valores passados da série ajudam na previsão de elementos futuros da série, que é uma característica essencial para previsões usando modelos lineares. O maior valor possível para previsões usando AR é o AR53 com valor de MSE 114,47 e MAD 8,78 apresentado em 4.6.

Tendo a Tabela 4.3 que mostram os resultados das previsões feitas usando modelos ARI com a frequência de 12 meses, é possível notar um ponto ótimo nos valores de previsão da série, que é entre o AR 10 e o AR 30 pois é onde encontrasse os menores valores MSE e MAD das previsões usando modelo AR. Foi encontrado o AR 23 como sendo o melhor modelo AR possível dentre os AR testados (entre AR1 até AR40) os demais modelos usando valores de p maiores provavelmente apresentam valores MSE maiores.

Retirando a frequência *airpassengers*

Por conta da tendência de subida apresentada na série estudada, era esperada que a retirada da diferenciação na modelagem fosse influenciar positivamente nos resultados. Porém quando se compara o método AR e o ARI os valores dos resultados da tabela ARI

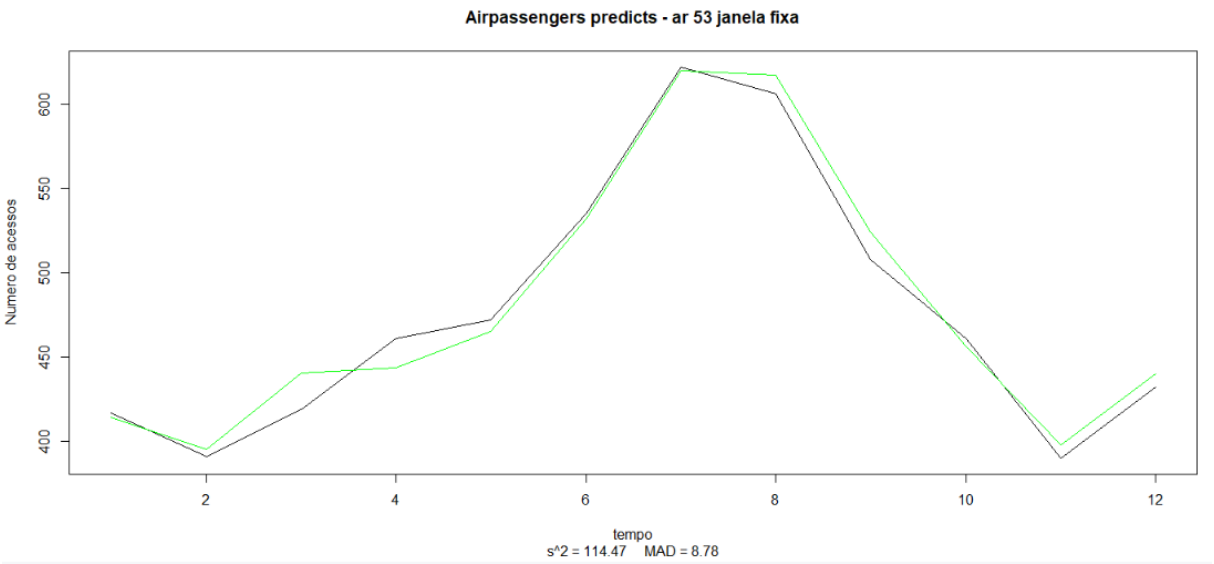

Figura 4.6: Modelo AR 53 previsão de Airpassengers.

Tabela 4.3: Resultados ARI AirPassengers

	MSE	MAS	MPE
ARI1	444,80	16,21	3,65
ARI2	472,40	16,67	3,76
ARI3	423,85	16,10	3,62
ARI5	417,27	16,14	3,64
ARI10	322,66	14,67	3,03
ARI20	251,52	12,34	2,68
ARI23	174,33	10,65	2,23
ARI40	222,90	12,03	2,63
ARI50	249,32	12,02	2,58
ARI60	532,47	20,23	4,42

foram piores na média, para esse conjunto de dados. Isso pode ser justificada devido ao uso da frequência na previsão.

Para exemplificar como a mudança na diferenciação e a retirada da frequência pode impactar nos resultados é necessário comparar dois modelos, um deles sendo o AR(10) mostrado na Figura 4.7 e o outro sendo o ARI(10) mostrado na Figura 4.8, porém dessa vez não será utilizado a frequência.

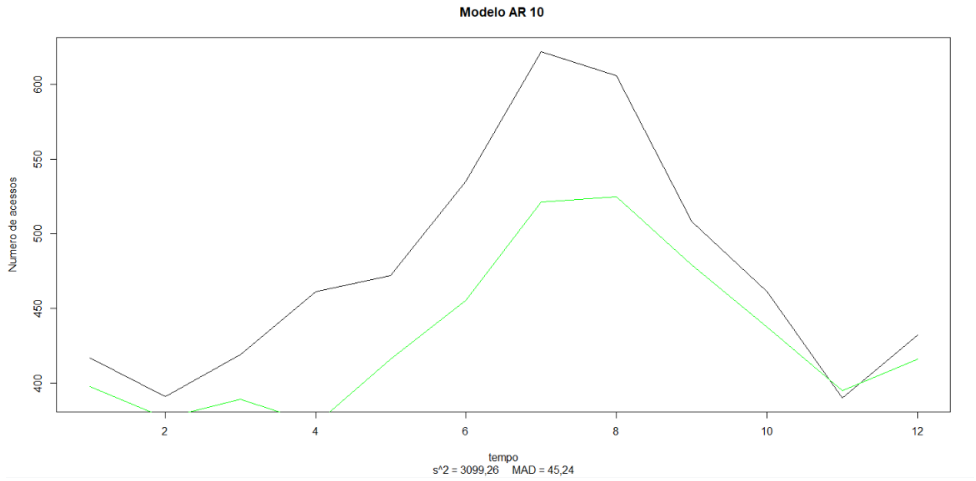

Figura 4.7: Modelo AR 10 sem frequência previsão de passageiros aéreos.

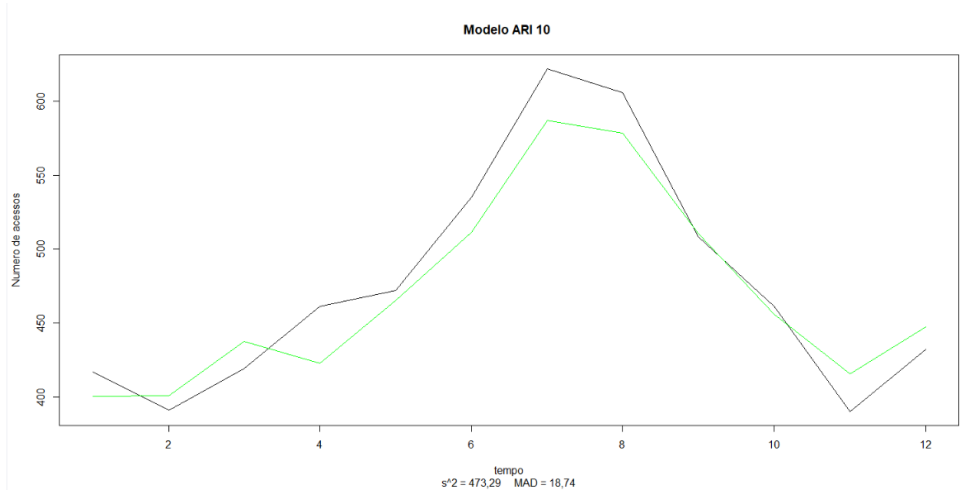

Figura 4.8: Modelo ARI 10 sem frequência previsão de passageiros aéreos.

A Tabela 4.4 mostra os resultados dos 4 modelos AR resumidos. O erro na parte no modelo AR sem frequência é muito maior quando comparado ao modelo com frequência, justamente pela força do modelo sazonal para essa série de tempo, quando essa componente é retirada é necessário um número grande de lags para reestruturar a previsão e resgatar as correlações. Quando se usa a diferenciação no modelo sem frequência a tendência de subida apresentada na série como um todo é adicionada, o que melhora em muito as previsões, observe a Figura 4.9

Tabela 4.4: Resumo de Resultados AR com e sem frequência

		MSE	MAS	MPE
com frequência	ARI10	348,59	14,37	3,23
com frequência	AR10	327,03	13,82	3,11
sem frequência	ARI10	473,29	18,74	3,91
sem frequência	AR10	3099,26	45,24	8,89

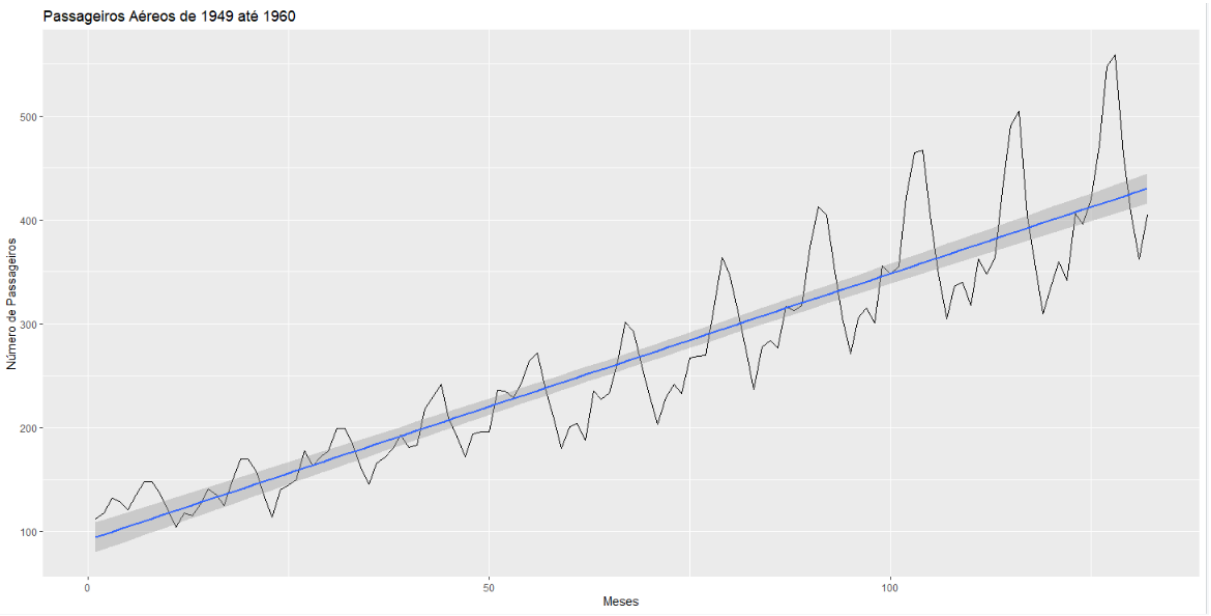

Figura 4.9: Tendência da série de tempo do número de passageiros aéreos.

4.2 Previsão do Número de Acessos

Usando os dados do número de acessos da página da Wikipédia do Netflix [21], uma variedade de modelos AR, MA, ARMA e ARIMA foram construídos. Alguns dos cálculos foram feitos utilizando Excell para facilitar a visualização dos cálculos e outros resultados foram adquiridos utilizando o Rstudio, devido a dificuldade computacional para cálculos mais complexos.

4.2.1 Explicando o Modelo do Artigo

Motivado em estudar um conjunto de dados bem conhecido e com uma sazonalidade não tão óbvia, foi escolhido a série de tempo fornecida pelo Google que registra a cada dia o número de acessos a página da Wikipédia. Das várias páginas disponíveis no estudo foi escolhido a do Netflix para efeitos de comparação por causa do estudo feito por Junyan Shao que comparou modelagem do ARIMA e o LSTM usando os mesmos dados [6].

O trabalho de Junyan e sua equipe apresenta um modelo ARI com valor p igual a 10, esse modelo apresenta um valor de MSE de 128,731 no artigo. Replicando os dados do estudo em Rstudio temos os resultados apresentados na Figura 4.10, onde conseguimos valores similares com MSE igual a 130,4. Interessante observar também o gráfico da Figura 4.11, que mostra em azul a modelagem feita e em preto os dados da série de tempo, onde podemos perceber uma similaridade grande da previsão com os dados reais.

```
arima(x = Netflix_zh[1:803], order = c(10, 1, 0), seasonal = list(order = c(0,
    0, 0), period = 7))

Coefficients:
          ar1      ar2      ar3      ar4      ar5      ar6      ar7      ar8
      -0.5144  -0.4897  -0.4163  -0.3726  -0.3291  -0.2670  -0.1691  -0.2003
s.e.   0.0352   0.0394   0.0424   0.0445   0.0454   0.0454   0.0444   0.0423
          ar9     ar10
      -0.1197  -0.0835
s.e.   0.0393   0.0352

sigma^2 estimated as 130.4:  log likelihood = -3091.69,  aic = 6205.38
```

Figura 4.10: Modelo ARI(10) proposto por Junyan Shao para os acessos as páginas do Wikipédia [6].

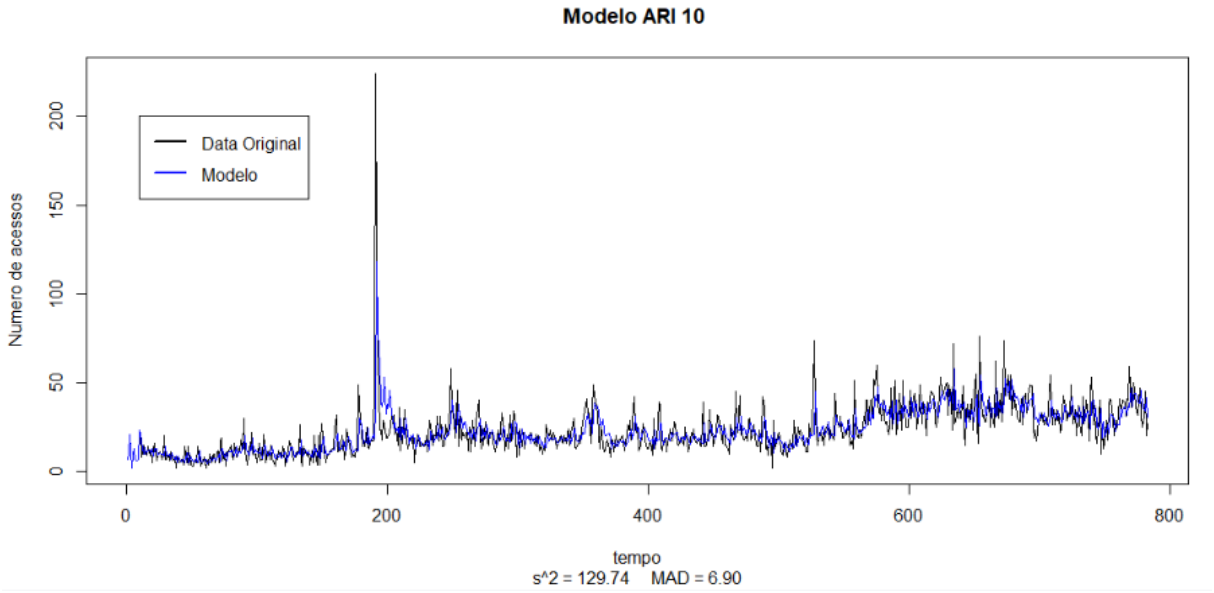

Figura 4.11: Gráfico do Modelo ARI(10) proposto por Junyan Shao [6].

No estudo de Junyan é comentado a suspeita de que um modelo ARI com p igual a 100 talvez obtenha valores MSE menores, configurando assim um modelo mais próximo do ideal, mas devido a falta de poder computacional não foi feito. Com o Rstudio esse estudo é possível e o modelo apresenta um valor de MSE de 113,3 mostrando também os coeficientes na Figura 4.12.

```
arima(x = Netflix_zh[1:803], order = c(100, 1, 0), seasonal = list(order = c(0,
    0, 0), period = 7))

Coefficients:
          ar1      ar2      ar3      ar4      ar5      ar6      ar7      ar8
      -0.5637  -0.5447  -0.5106  -0.5118  -0.4875  -0.4484  -0.3682  -0.4400
s.e.   0.0353   0.0406   0.0449   0.0485   0.0518   0.0546   0.0568   0.0583
          ar9     ar10     ar11     ar12     ar13     ar14     ar15     ar16
      -0.4045  -0.3764  -0.3833  -0.3045  -0.2380  -0.2523  -0.2536  -0.2076
s.e.   0.0603   0.0620   0.0634   0.0650   0.0659   0.0664   0.0670   0.0676
         ar17     ar18     ar19     ar20     ar21     ar22     ar23     ar24
      -0.2508  -0.2136  -0.2485  -0.2364  -0.2093  -0.1062  -0.1891  -0.2269
s.e.   0.0680   0.0686   0.0690   0.0695   0.0700   0.0704   0.0704   0.0706
         ar25     ar26     ar27     ar28     ar29     ar30     ar31     ar32
      -0.2231  -0.1966  -0.1806  -0.2128  -0.2227  -0.1975  -0.1088  -0.1471
s.e.   0.0710   0.0714   0.0717   0.0719   0.0722   0.0725   0.0728   0.0728
         ar33     ar34     ar35     ar36     ar37     ar38     ar39     ar40
      -0.1628  -0.1848  -0.1759  -0.1442  -0.1765  -0.2043  -0.1561  -0.2011
s.e.   0.0729   0.0731   0.0734   0.0738   0.0740   0.0743   0.0747   0.0749
         ar41     ar42     ar43     ar44     ar45     ar46     ar47     ar48
      -0.1386  -0.1231  -0.1734  -0.1895  -0.1326  -0.1085  -0.0875  -0.1101
s.e.   0.0753   0.0755   0.0756   0.0759   0.0762   0.0762   0.0763   0.0762
         ar49     ar50     ar51     ar52     ar53     ar54     ar55     ar56
      -0.0949  -0.0552  -0.0720  -0.1186  -0.1321  -0.0649  -0.0856  -0.0783
s.e.   0.0761   0.0761   0.0761   0.0760   0.0760   0.0761   0.0760   0.0758
         ar57     ar58     ar59     ar60     ar61     ar62     ar63     ar64     ar65
      -0.0402   0.0056   0.0594   0.0069   0.0108   0.0413   0.023  -0.0075   0.0388
s.e.   0.0755   0.0753   0.0751   0.0749   0.0746   0.0744   0.074   0.0737   0.0735
         ar66     ar67     ar68     ar69     ar70     ar71     ar72     ar73
       0.0413  -0.0052  -0.0184  -0.0470  -0.0184  -0.0648  -0.1031  -0.0828
s.e.   0.0732   0.0729   0.0727   0.0725   0.0724   0.0721   0.0717   0.0714
         ar74     ar75     ar76     ar77     ar78     ar79     ar80     ar81
      -0.0468  -0.1054  -0.0879  -0.0971  -0.074  -0.0920  -0.0470  -0.0825
s.e.   0.0712   0.0708   0.0706   0.0702   0.070   0.0699   0.0695   0.0691
         ar82     ar83     ar84     ar85     ar86     ar87     ar88     ar89
      -0.0542  -0.0532  -0.0678  -0.0127  -0.0513  -0.0541  -0.0373  -0.0194
s.e.   0.0685   0.0681   0.0675   0.0671   0.0664   0.0658   0.0653   0.0644
         ar90     ar91     ar92     ar93     ar94     ar95     ar96     ar97
      -0.0619  -0.0262  -0.0379  -0.0446  -0.0648  -0.0409  -0.0337   0.0069
s.e.   0.0628   0.0614   0.0596   0.0575   0.0561   0.0537   0.0509   0.0477
         ar98     ar99    ar100
      -0.0490  -0.0203  -0.0352
s.e.   0.0443   0.0402   0.0354

sigma^2 estimated as 113.3:  log likelihood = -3037.98,  aic = 6277.95
```

Figura 4.12: Valores dos coeficientes ar100 para número de acessos Netflix.

Observando o gráfico de parte do modelo do ARI(100) apresentado na Figura 4.13, é possível comparar com o gráfico do ARI(10) e notar que a MAD entre os dois modelos também diminui, o que é um indicativo de uma uma modelagem mais bem feita.

O problema desses modelos é que por mais que tentemos melhorar e aproximar a modelagem dos valores reais observados na série de tempo, não se pode observar uma aplicação desse modelo em um cenário real, uma vez que nenhuma previsão é feita de fato, o que ocorre é um estudo dos modelos e sua eficiência na teoria. Por esse motivo na próxima seção, algumas alterações foram feitas na série de tempo com o objetivo avaliar as previsões que podem ser feitas com ARIMA, tentando avaliar sua eficiência em um cenário real ligado ao comportamento de acesso de usuários.

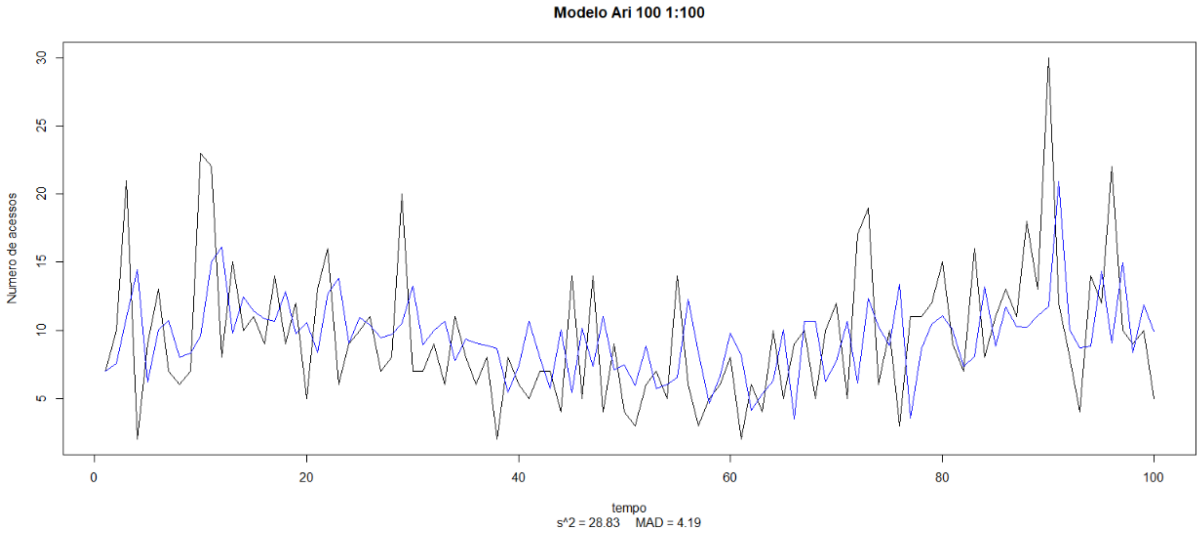

Figura 4.13: Corte do gráfico do modelo ARI(100) mostrando de 1 até 100 dos 803 elementos usados.

4.2.2 Avaliando os dados

Para acontecer um previsão com os dados do acesso da página da Wikipedia do Netflix foram retirados os últimos 20 valores da série de tempo e as previsões foram feitas para tentar melhor se aproximar desses valores.

Com a avaliação do gráficos ACF e PACF apresentados nas Figuras 4.14 e 4.15, temos um modelo inicial como sendo o AR(1) ou o AR(10). O que acontece é que algumas correlações podem não ser levadas em consideração, uma vez que o conjunto de dados é muito longo e dificulta a visualização das correlações usando ACF e PACF.

Quanto maior o número de p e q usados no modelo maior o número de elementos passados usados como referência e maior o número de erros passados utilizados para a previsão, o que acarreta na suposição de que os valores passados da série podem ser usados para prever valores futuros da mesma série. Então quando o aumento no valor de lags usados para a previsão acontece, assumimos também que a série se comporta de maneira parecida até p valores.

Reforçar a suposição inicial do modelo ARIMA é importante para justificar alguns resultados. Um estudo dos modelos AR, MA, ARMA e ARIMA é feita mostrando além dos resultados para a previsão dos últimos 20 valores da série tempo, mas também do modelo feito para comparação com os resultados apresentados na Seção 4.2.1.

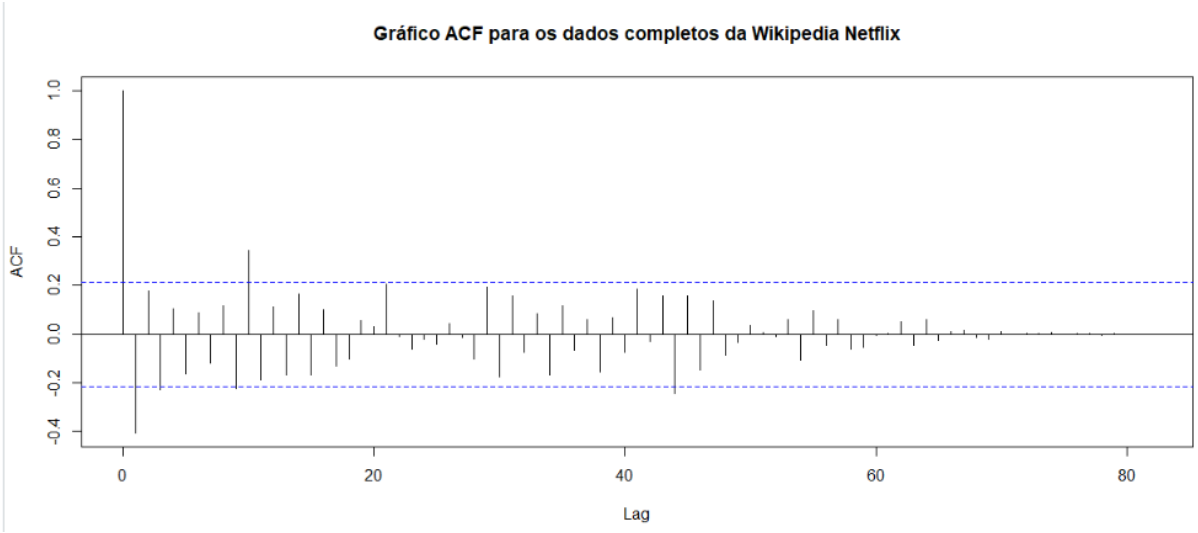

Figura 4.14: Avaliação ACF número de acessos a página Wikipédia Netflix.

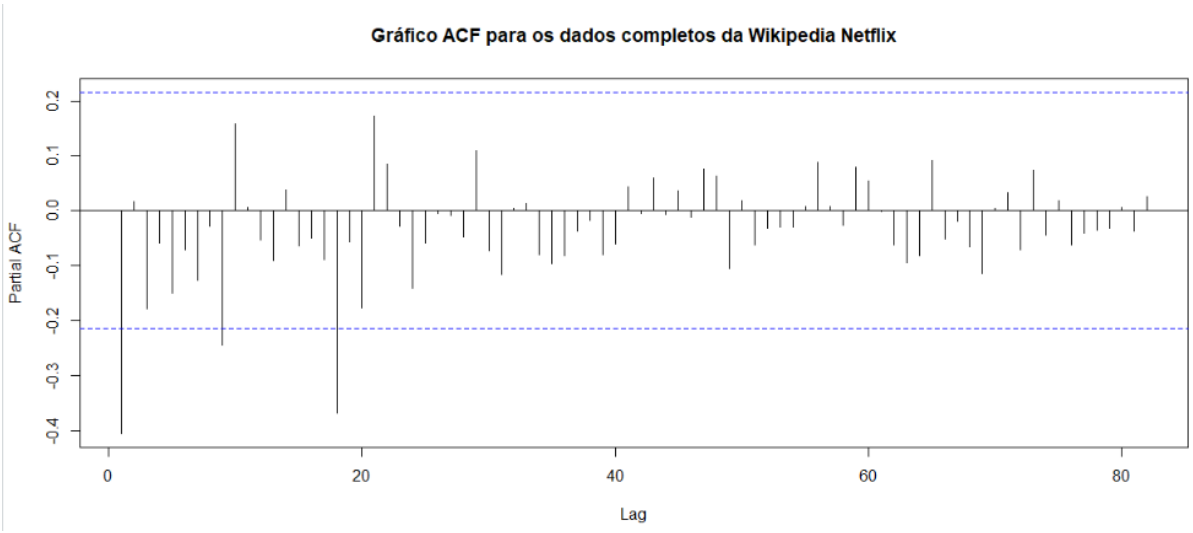

Figura 4.15: Avaliação PACF número de acessos a página Wikipédia Netflix.

4.2.3 Resultados prevendo os últimos 20

ACF e PACF

Inicialmente para a construção do modelo ARIMA é feita a avaliação dos gráficos ACF e PACF que são mostrados nas Figuras 4.16 e 4.17. Por serem quase os mesmos dados da seção anterior o modelo inicial também é parecido, um AR(1) ou AR(10).

O gráfico da previsão feita com esse modelo é mostrado na Figura 4.18, onde é possível perceber um certo decaimento com o passar das previsões. Após a previsão número 10 a variação dos valores previstos são muito menores que os primeiros valores previsto, esse

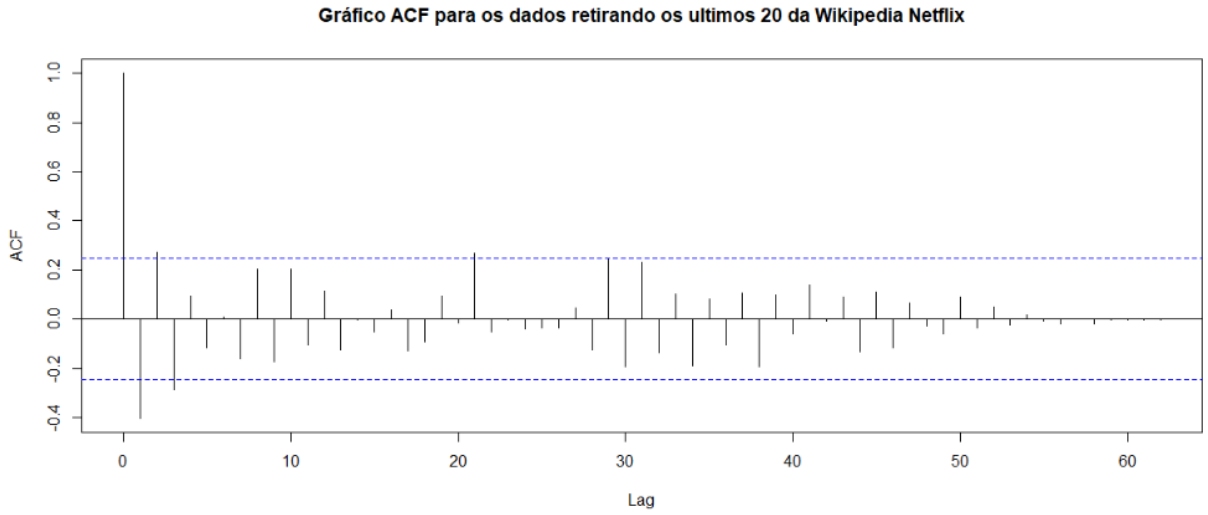

Figura 4.16: Avaliação ACF número de acessos a página Wikipédia Netflix.

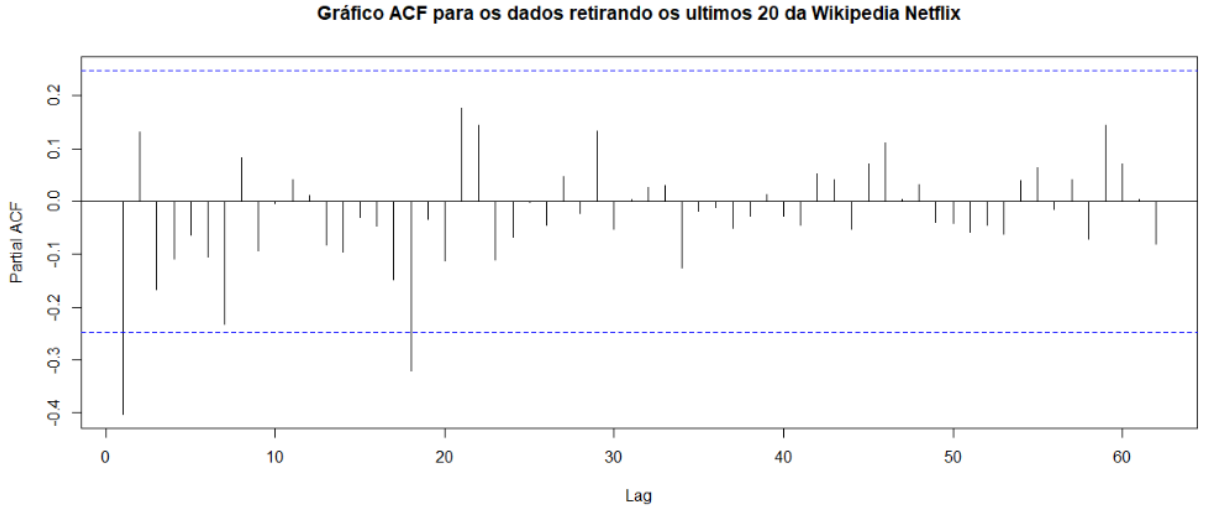

Figura 4.17: Avaliação PACF número de acessos a página Wikipédia Netflix.

decaimento na variação não é um fator positivo. As previsões feitas com modelos ARIMA com p e q menores que o número de previsões, tendem a voltar para um valor médio, ou seja, para melhorar as previsões os valores p e q tem que ser maiores que o número de elementos que se quer prever.

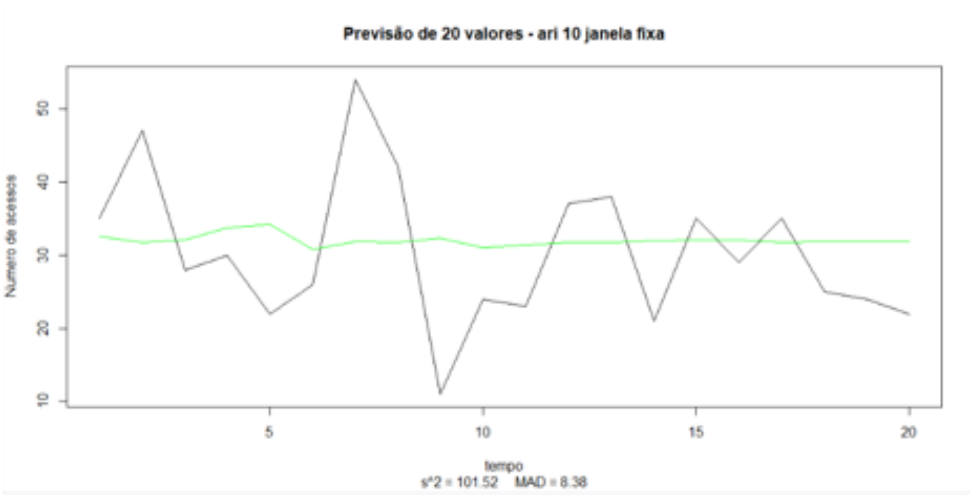

Figura 4.18: Previsão com 10 lags para o número de acessos netflix - Avaliação ACF e PACF.

No caso proposto se deseja prever 20 valores, então os valores p e q tem que ser maiores ou iguais a 20. Na Figura 4.20 é mostrado um exemplo com um gráfico onde os valores p e q são respectivamente 5 e 20, e o decaimento na variação da série não acontece tão rapidamente.

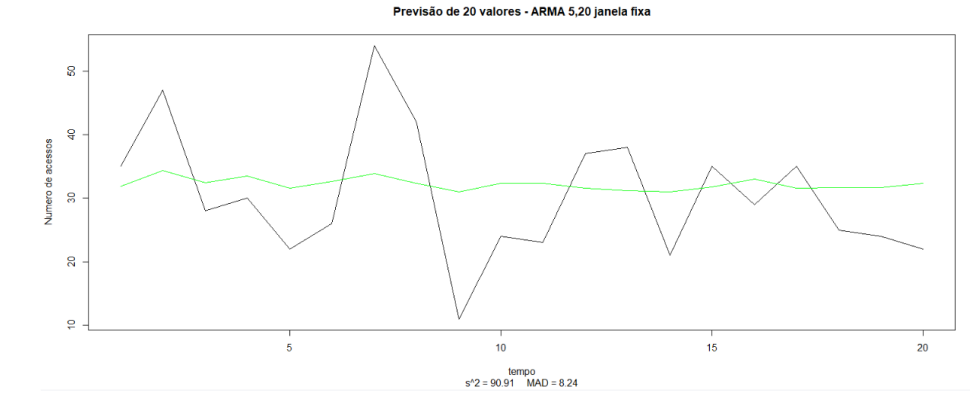

Figura 4.19: Resultado do modelo ARMA(5,20) para o número de acessos à página do netflix.

Tabela 4.5: Resultados AR Netflix

	MSE	MAS	MPE
AR5	124,98	8,01	25,43
AR10	101,43	7,44	25,76
AR20	96,98	8,38	34,11
AR40	97,41	8,23	33,84
AR50	95,38	8,22	32,73
AR60	92,85	8,06	31,95
AR70	89,52	7,96	31,55
AR80	96,48	8,24	32,37
AR100	96,29	8,37	33,41
AR200	118,93	8,55	37,64
AR210	132,97	8,56	38,23
AR350	421,65	17,57	67,23

AR MA ARMA ARIMA - valores acima de 20

Apesar de usar valores p e q superiores a 20 o modelo pode não estar otimizado, por isso foi feito em uma tabela mostrando os resultados MSE e MAD de varios modelos.

Os melhores resultados da previsão AR está por volta do valor AR 100 onde as previsões não apresentam melhoras significativas na Tabela 4.5. Também é mostrado a qualidade do modelo em prever os valores do próprio modelo esses valores de MSE são dados na modelagem e quanto maior o número de p um modelo AR tiver melhor será a modelagem. Essa característica não resulta em uma melhora para as previsões, é possível verificar isso apartir do AR(200) onde a modelagem melhora com relação a AR(100) porém as previsões pioram, os valores de MAD e MSE aumentam.

O nome dado para esse fenômeno é *overfitting*, que pode ser descrito no contexto de modelagem de séries temporais como sendo um erro na escolha dos parâmetros, onde existe um excesso na caracterização do modelo [22]. Isso pode acarretar em uma modelagem complexa, o modelo construído é útil apenas para ser usado em um único contexto e por isso não é algo desejável para previsões.

A detecção de *overfitting* pode ser complexa e por isso não é feito nenhum método para evitar que ele ocorra mas de maneira experimental avaliando o MSE do modelo é possível ter uma ideia de onde isso pode estar acontecendo. Mais um exemplo de *overfitting* pode ser visto na Tabela 4.7, em algum modelo entre o MA 40 e o MA 100 o valor de MSE das previsões começa a subir e o valor de MSE do modelo continua a descer.

Quando misturamos *moving average* e *autoregression* usando ARMA temos uma variedade grande de possibilidades de combinações, com o objetivo de simplificar o agrupamento dos dados foram feitos apenas alguns deles.

Observando a tendência dos dados pode-se dizer de antecipadamente que não existe

Tabela 4.6: Qualidade do Modelo AR Netflix

	MSE	MAS	MPE
AR5	132,92	7,15	43,22
AR10	128,67	7,01	41,36
AR20	124,52	6,83	39,52
AR40	120,72	6,70	38,74
AR50	118,83	6,63	38,16
AR60	116,84	6,56	37,68
AR70	113,25	6,22	32,39
AR80	111,48	6,14	31,27
AR100	109,08	5,96	29,70
AR200	45,64	4,36	19,92
AR210	43,51	4,22	19,20
AR350	14,92	2,21	8,79

Tabela 4.7: Resultados MA Netflix

	MSE	MAS	MPE
MA5	142,37	8,71	26,88
MA10	137,49	8,54	26,29
MA20	131,46	8,28	25,85
MA40	87,87	7,64	28,97
MA100	111,70	7,95	29,79

Tabela 4.8: Qualidade do modelo MA Netflix

	MSE	MAS	MPE
MA5	140,16	7,61	49,31
MA10	134,31	7,37	46,55
MA20	128,47	7,12	43,72
ma40	121,84	6,92	41,76
MA100	110,32	6,53	39,14

Tabela 4.9: Resultados ARMA Netflix

	MSE	MAS	MPE
ARMA5,5	102,33	8,58	35,43
ARMA10,5	101,17	8,45	34,62
ARMA20,5	93,06	8,28	33,07
ARMA5,10	101,57	8,47	35,07
ARMA5,20	90,91	8,24	33,44
ARMA10,10	105,30	8,54	35,60
ARMA10,20	96,84	8,31	34,38
ARMA20,20	123,34	9,36	39,06
ARMA100,100	557,91	19,99	79,31
ARMA350,100	516,19	19,05	70,63
ARMA350,200	344,98	15,27	59,06

Tabela 4.10: Qualidade do Modelo ARMA Netflix

	MSE	MAS	MPE
ARMA5,5	124,93	6,75	38,75
ARMA10,5	124,26	6,70	38,16
ARMA20,5	122,27	6,77	39,44
ARMA5,10	123,86	6,71	38,43
ARMA5,20	122,49	6,74	38,42
ARMA10,10	121,69	6,71	38,37
ARMA10,20	121,28	6,70	38,42
ARMA20,20	121,69	6,71	38,37
ARMA100,100	73,63	5,26	26,38
ARMA350,100	7,45	1,6	6,71
ARMA350,200	4,33	1,21	5,25

Tabela 4.11: Resultados ARIMA Netflix

	MSE	MAS	MPE
ARIMA1,1	112,81	8,15	34,77
ARIMA5,5	108,04	8,76	37,09
ARIMA10,5	106,69	8,77	37,04
ARIMA20,5	98,72	8,49	35,45
ARIMA5,10	114,84	8,96	38,10
ARIMA5,20	125,97	9,42	39,07
ARIMA10,10	116,04	8,88	38,39
ARIMA10,20	103,13	8,55	36,25
ARIMA20,10	100,25	8,37	35,29
ARIMA20,20	111,53	8,94	37,89

uma componente de tendência muito grande, e por esse motivo os resultados do modelo ARIMA mostrados a Tabela 4.11 podem não mostrar uma melhora comparado com o modelo ARMA.

Tabela 4.12: Resultados ARIMA Netflix

	MSE	MAS	MPE
ARIMA1,1	126,12	6,70	36,83
ARIMA5,5	123,54	6,67	36,83
ARIMA10,5	122,95	6,70	37,01
ARIMA20,5	121,84	6,69	36,79
ARIMA5,10	123,66	6,72	37,09
ARIMA5,20	120,75	6,70	37,05
ARIMA10,10	120,55	6,67	37,01
ARIMA10,20	120,80	6,63	36,69
ARIMA20,10	118,76	6,60	35,54
ARIMA20,20	108,11	6,43	34,53

Os melhores resultados entre todos esses apresentados na Tabela 4.9 são os modelos ARMA(5,20) e MA(40) que são mostrados respectivamente na Figura 4.20 e na Figura 4.21, onde é possível notar correlações interessantes.

No modelo ARMA(5,20) a variação das previsões são mais amenas quando comparadas com a do modelo MA(40), apresentando resultados médios em torno de 30 acessos. Quando observamos o modelo MA(40) além de ser possível observar uma boa captura de correlações nas previsões 3 até a 5 e também na porção de 15 até 19, os valores médios de MSE são os menores entre os apresentados.

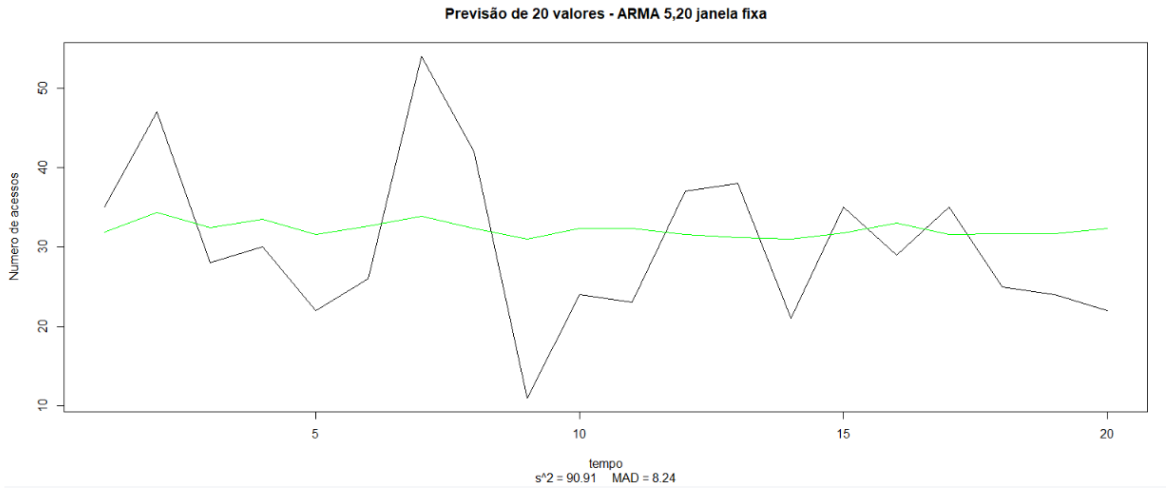

Figura 4.20: Previsões do modelo ARMA(5,20) número de acessos à página Netflix.

Com os resultados adquiridos dos diversos modelos construídos algumas conclusões podem ser feitas. Apesar das dificuldades por causa da falta de sazonalidade ainda é possível capturar algumas características da série usando modelos lineares.

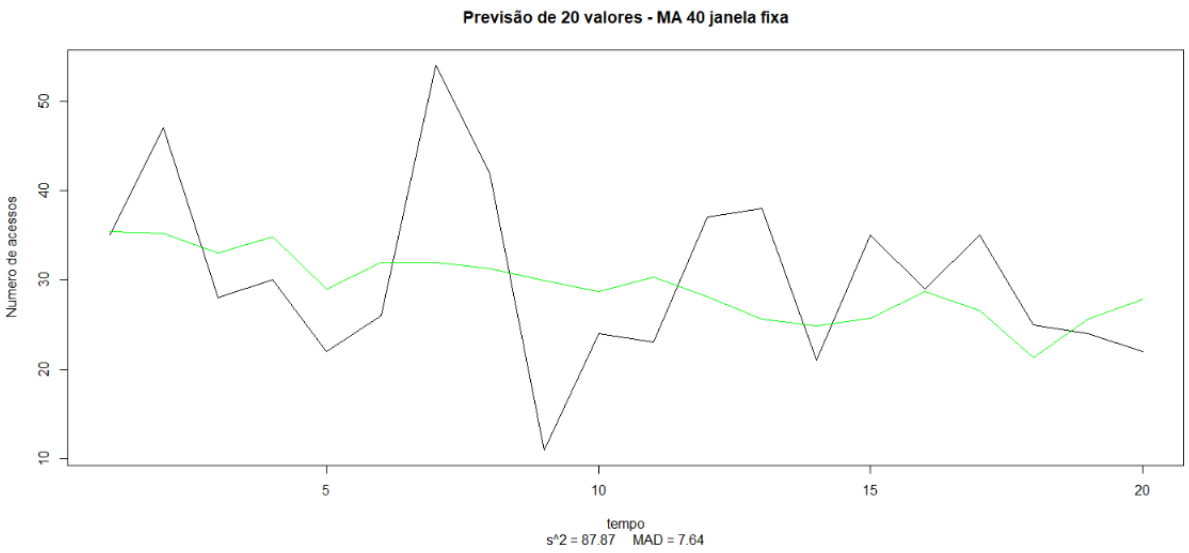

Figura 4.21: Previsões do modelo MA(40) número de acessos à página Netflix.

Foram feitos mais alguns testes para tentar achar um melhor modelo ARMA possível, esse modelo é o ARMA(57,55) que tem MSE de 70,64 e MAD 6,64, mostrado na Figura 4.22. Valores superiores a esses apresentam somente uma piora nos resultados MSE.

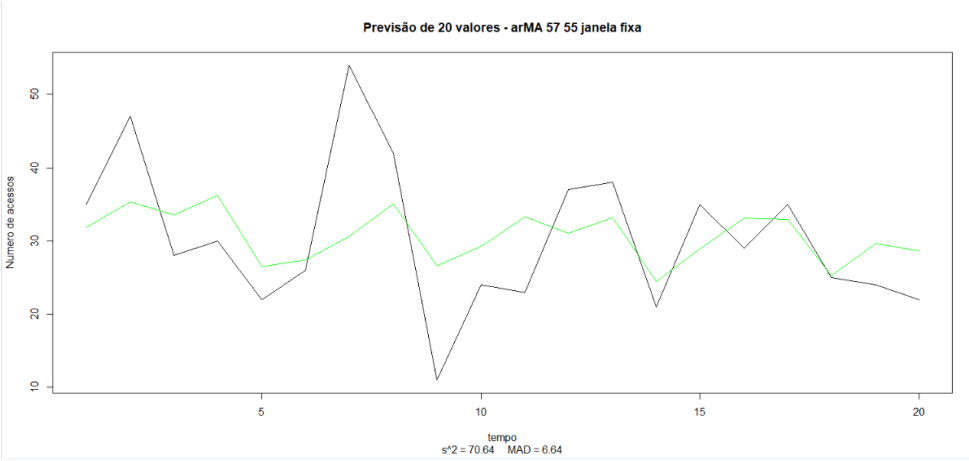

Figura 4.22: Previsões do modelo ARMA(57,55) número de acessos à página Netflix.

4.3 Previsões dos valores de potência

Para prever os valores de uma série de tempo existem uma variedade de variáveis que influenciam nos resultados. Nesse capítulo estudamos uma diversidade de modelos ARIMA

com diferentes entradas e parâmetros tentando, compreender se seria possível utilizar de modelos lineares para prever o comportamento de usuário, no contexto de redes.

Durante o capítulo ocorre um estreitamento no escopo da pesquisa, indo de um cenário mais geral e difícil de se prever, para um cenário mais específico, tentando se aproximar cada vez mais dos parâmetros ótimos para previsões usando métodos lineares. Utilizando da experiência adquirida nas seções passadas o conjunto de dados inicial é alterado inicialmente, justamente para se encaixar num perfil mais periódico. Em seguida uma diversidade de variáveis que são importantes no resultado da previsão são modificadas, entre elas os valores de p, q e d dos modelos ARIMA, a janela de previsão, os valores de frequência na série de tempo.

Com o objetivo inicial de identificar o modelo ideal para a previsão da série de tempo e de justificar algumas decisões tomadas no estudo do mesmo, foi necessário um diagnóstico das características da série, logo de imediato percebe-se que existe uma grande tendência para os valores de potência, a maior parte deles é normalizado e varia entre 1 e 2, que equivalem ao canal do transmissor estando vazio ou em IDLE, ou seja, não temos nenhum nó efetivamente transmitindo naquele espaço de tempo. Outros valores da série são das transmissões, que variam bastante porém com uma tendência para valores normalizados de 50 até 150. Com relação a ordenação entre os números vemos pela Figura 4.23 que existe um padrão comum entre os valores, justificado pelo próprio padrão CSMA/CA, onde temos um tempo de espera onde o canal fica vazio, um tempo de transmissão, seguido de uma espera pela resposta do receptor e depois uma pequena transmissão do receptor (ACK).

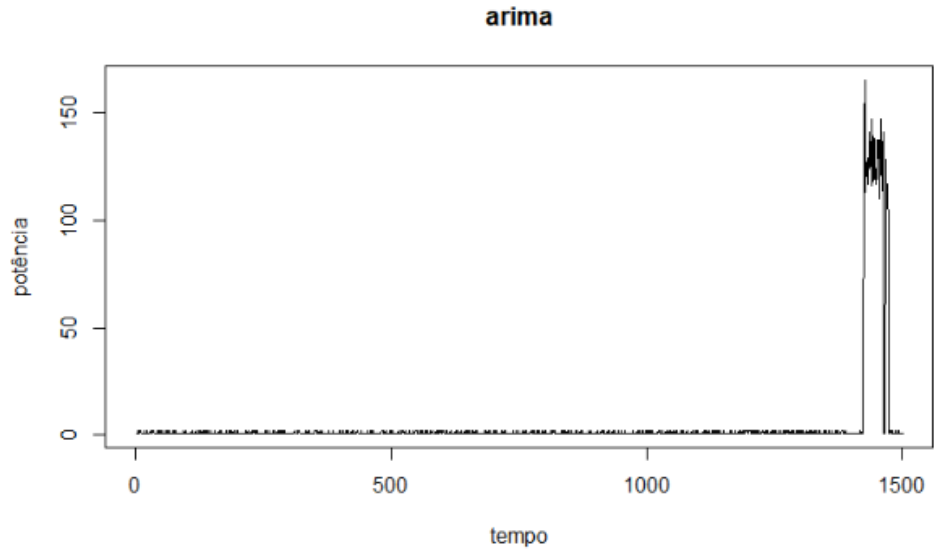

Figura 4.23: Exemplo de gráfico de Energia com tempo ocioso e transmissão.

A informação mais importante retirada dos dados iniciais é que o ciclo para a categorização de uma transmissão IDLE/DATA/SIFS/ACK, mostrado na Figura 4.23. Uma previsão com modelos lineares teria dificuldades em fazer previsões nessa serie de tempo. Um modelos ARIMA(3,1,3) leva em consideração os 3 valores passados da série do tempo para influenciar as previsões futuras, dessa forma um modelo que resgataria inicialmente sua previsão com os valores de amostra 1 e 2 de um canal em idle por exemplo, precisaria de um valor de p ou q altos, montando um modelo AR(5000) para considerar uma transmissão.

O resultado da insistência na escolha de se trabalhar com os valores puros de potência resulta em um impedimento computacional, uma vez que se faz necessário escolha de um modelo com valores de lags muito altos,um ARIMA(500,1,1000) por exemplo, essa escolha faz com que seja necessário um cálculo que toma muito tempo e assim sendo inapropriado para utilização no contexto da identificação da oportunidade.

Outro fator que influência negativamente e dificulta a previsão em séries muito longas é a perda de precisão da série. Não é incomum que após alguns valores de previsão a série perca vigor, aumentando a taxa de erro ou até dando valores de previsão fixos. Isso acontecia bastante nesse caso, o que era esperado uma vez que grande parte dos valores da série são entre 1 e 2 a previsão provavelmente irá estar entre esses valores.

Em outras palavras, temos que os dados puros obtidos pelos valores de potência não são os ideais para a previsão utilizando o modelo ARIMA, uma vez que a escolha para os valores dos lags depende da correlação do valor atual da série com um valor passado.

4.3.1 Modelagem da série de tempo

Apesar da dificuldade inicial na configuração do modelo a alteração e remodelagem dos dados faz parte do processo de aprimoramento da previsão, com isso em mente algumas mudanças na série podem ser feitas para se obter um resultado mais agradável.

A alteração do valor de amostragem foi considerada, uma vez que isso diminuiria a distância entre um valor de IDLE e de transmissão, facilitando a identificação de correlação entre os valores da série diminuindo o tempo necessário para cálculo da previsão. O problema dessa abordagem se dá no ponto em que as amostras iniciais já estavam espaçadas com um tempo alto de captura, um aumento desse valor de amostragem poderia fazer com que um valor de transmissão ficasse curto demais ou que até se perdesse a captura de um ACK dentro da transmissão.

Outra solução considerada foi a de ao invés de utlizar os valores puros da potência, passar a usar o número de amostras consideradas IDLE e o número de amostras consideradas DATA, esperando obter uma série com frequência 4 de IDLE/DATA/SIFS/ACK.

Alguns resultados positivos foram percebidos nessa abordagem, o ponto principal a ser notado aqui é que as previsões em sua maioria tinham os valores de SIFS e ACK iguais, isso é o esperado uma vez que na definição do CSMA/CA o tempo dessas etapas é fixo, o passo mais lógico seguido foi o de retirar esses valores da previsão e trabalhar com somente os valores de IDLE da amostra de potência. Como os valores de transmissão também não era de grande interesse para a caracterização de uma oportunidade descartamos afim de facilitar a previsão e focar no essencial.

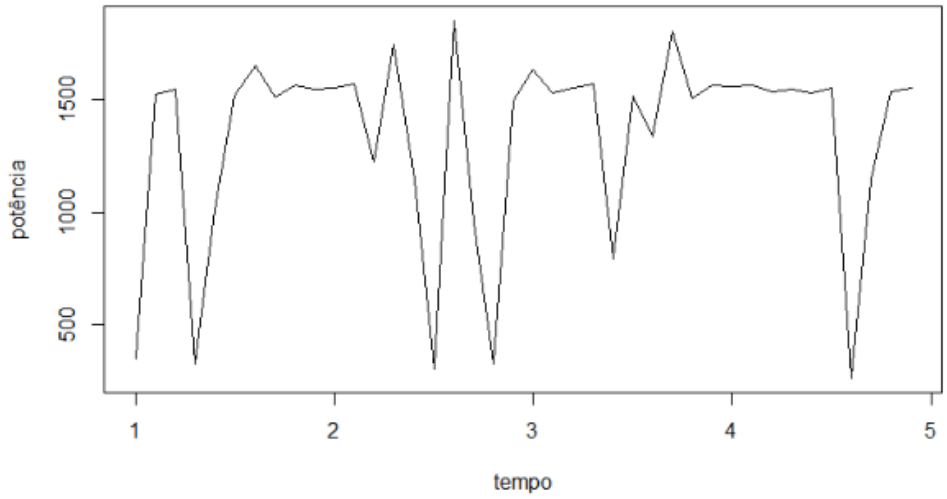

Figura 4.24: Gráfico do intervalo entre transmissões.

A série de tempo modificada com somente os valores de IDLE ficou com um total de 96 valores, cada um deles com o número de amostras de canal vazio obtidos da coleta de potência, verifique a Figura 4.24.

4.3.2 Modelagem do ARIMA

Seguindo as mesmas etapas da criação do modelo ARIMA feita usando o número de passageiros mostrada na Seção 3.2, foi iniciado a modelagem do modelo ARIMA para os dados apresentados na Figura 4.25. A primeira modelagem será feita da seguinte forma: dos 96 valores disponíveis, serão usados os 40 iniciais para a construção de uma pseudo série de tempo, modelos usando valores (p, d, q) diferentes vão ser montados e a sua performance avaliados. Na Figura 4.25 e 4.26 que mostra o gráfico ACF e PACF respectivamente para os primeiros 40 valores da série de tempo Figura 4.24

Os valores baixos de correlação(menores que 0,2) reveladas nos lags das imagens PACF e ACF mostram que para essa amostra de dados o modelo ARIMA talvez não seja o suficiente para prever os valores da série, nesses casos a série diferenciada ajuda a revelar

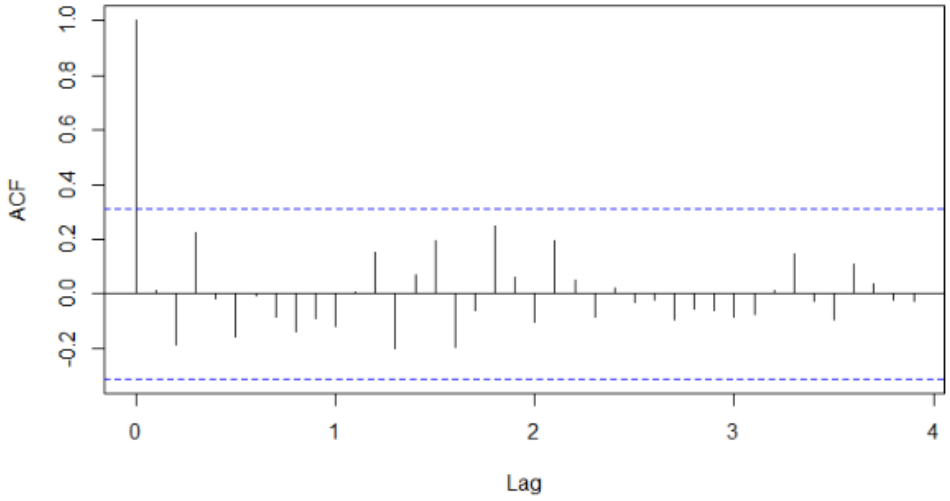

Figura 4.25: ACF com 40 lags, dados com beacon.

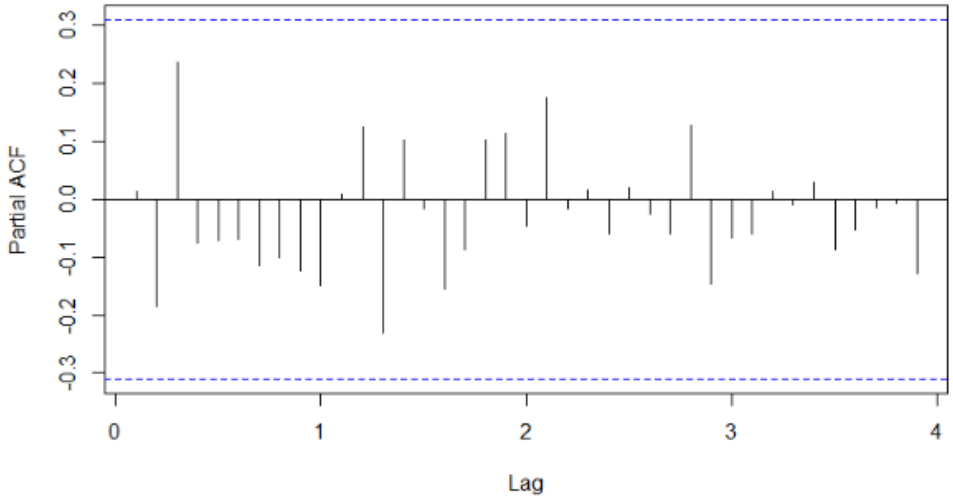

Figura 4.26: PACF com 40 lags, dados com *beacon*.

essa correlação entre os números, por isso, novos testes são feitos nas figuras 4.27 e 4.28. Os beacons representam transmissões que fazem controle da rede Wi-fi. No nosso estudo é interessante remover as transmissões atreladas aos beacons, uma vez que esses não fazem parte da caracterização de usuário primário proposta.

Como citado anteriormente um fator importante para a modelagem da série de tempo é a frequência ou sazonalidade da série, como no caso desse modelo não temos noção de alguma sazonalidade óbvia, o valor de frequência escolhido foi de 10 amostras. Esse número se dá pelo fato da quantidade de amostras ser limitado, um valor de frequência

muito baixo pode acarretar em imprecisões na previsão e números muito altos acabariam com a quantidade de amostras da série.

Os dados remodelados apresentam uma correlação que é mais facilmente percebida, o teste PACF da amostra em questão é mostrado na Figura 4.27, onde os lags mais significantes tem valor de 1, ou seja, para esses dados o valor de q do modelo ARIMA é também igual ou próximo a 1.

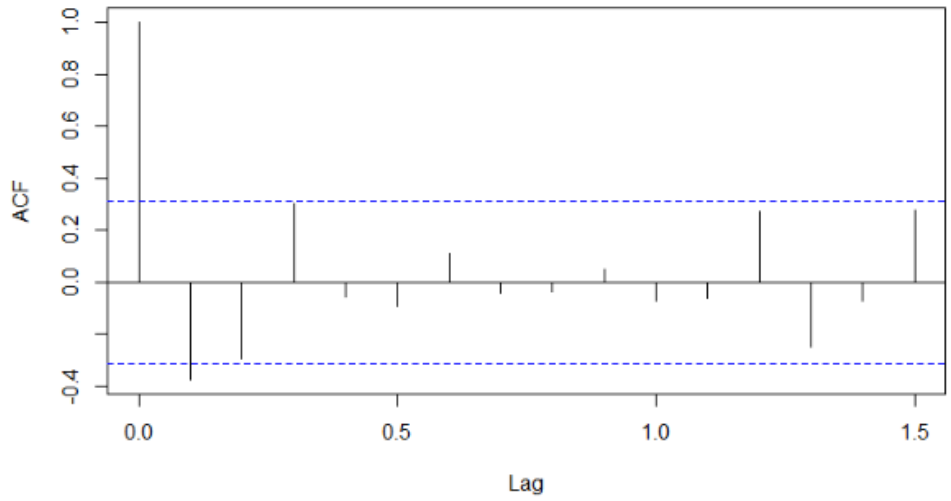

Figura 4.27: ACF com 40 lags, aplicando difereciação.

O teste PACF mostrado na Figura 4.28 revela que o número de lags q para esse exemplo é de 2, resultando em um ARIMA(2,1,1) para as primeiras 40 amostras. Com esse valores de lags e com a previsão dos próximos 40 valores da série, temos os resultados apresentados 4.29.

Devido ao valor escolhido para a frequência da série de 10 amostras, o modelo espera que exista uma repetição de tendências após 10 previsões, e por causa disso existe um padrão que também é visto nas previsões de 11 até 20 e novamente de 21 até 30 e assim por diante. Para os primeiros 10 valores de previsão é possível visualizar que o modelo apresenta uma taxa de erro médio menor quando comparado com a previsão da série como um todo, não só pela natural perda de precisão do ARIMA mas pela escolha do valor de frequência.

Existe a possibilidade de assumir que o modelo ARIMA não tenha sazonalidade, isso afetaria na previsão de modo a fazer os valores previstos tenderem para uma média, essa solução não é boa para o problema em questão, logo foi preferido escolher a abordagem com frequência. Para os 10 primeiros valores de previsão temos os resultados apresentados na Figura 4.29 em preto a amostra original, os valores previstos em verde, a taxa de erro

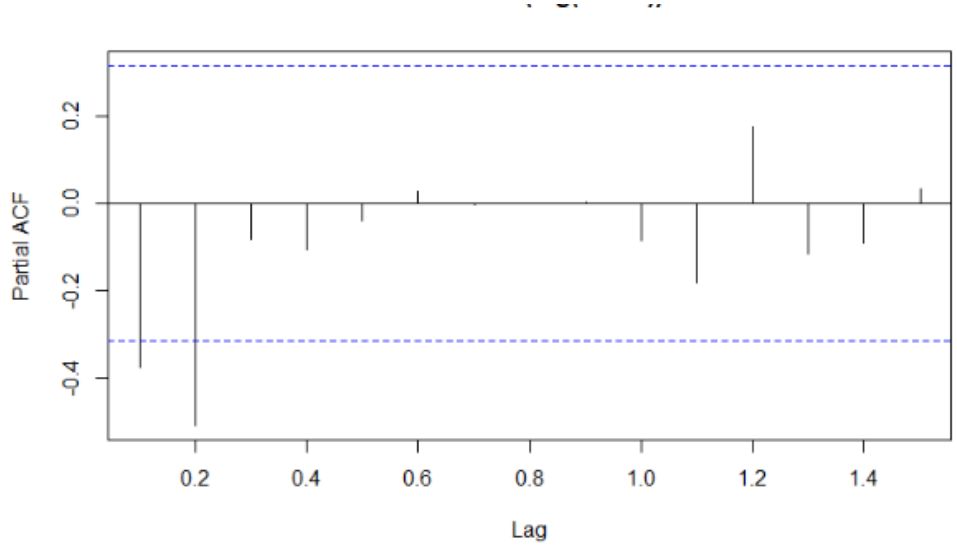

Figura 4.28: PACF com 40 lags, aplicando difereciação.

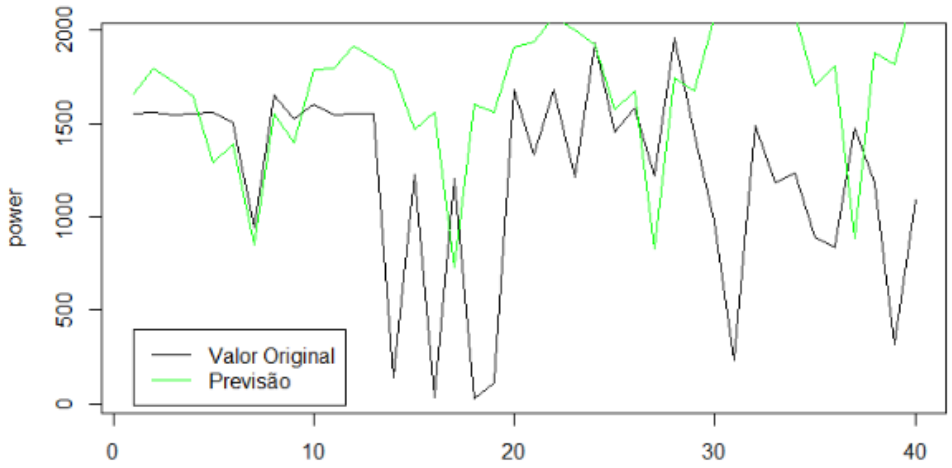

Figura 4.29: Previsão com sazonalidade 10 dados com beacon, ARIMA(2,1,1).

médio dessa previsão foi de 706,42 *slots* e com erro quadrático de 617078,50 os resultados apresentados na Tabela 4.13.

O problema da abordagem com sazonalidade está na alta taxa de desperdício, devido a tendência atribuída a série. Nesse caso foi atribuído uma constante positiva para as previsões e isso influenciou muito para esse o alto nível de desperdício. Para melhorar a previsão removemos a sazonalidade e verificamos os melhores modelos na próxima seção.

Tabela 4.13: Resultados variando número de previsões

	MSE	MAD	MPE	TC	TD
40 previsões	617078.5	589.36	358.78	706.42	75
10 previsões	29487.17	159.42	10.58	174.18	54.54

4.3.3 Frequência e modelo ARIMA

O modelo ARIMA sozinho, sem a utilização da frequência, tende a perder precisão no momento da previsão. As primeiras previsões feitas usando os mesmos dados conseguem prever alguns valores, a previsão depois de poucas previsões contém somente uma componente constante, o que resulta na previsão sendo linear, o que nesse caso não reflete a realidade. Para melhorar esse resultado avaliamos novamente a correlação entre os elementos da série usando o teste ACF. Verificamos que até o lag 21 ocorre uma correlação positiva, por esse motivo escolhemos esse valor para configurar o número de elementos a serem considerados para a nova previsão não sazonal.

Para as próximas previsões a série de tempo contém uma janela de 21 elementos. Um modelo ARIMA, AR, MA ou ARMA usado para gerar um valor de previsão, a seguir a janela se move para frente com o valor real da amostra e se faz uma nova previsão.

Avaliando inicialmente o modelo AR da previsão com o menor MAD é o AR 1, a Figura 4.30 mostra que quanto maior o número de elementos a serem levados em consideração para a previsão piorá o erro médio da série. Essa conclusão foi tirada devido a disparidade da curva observada com relação à previsão feita.

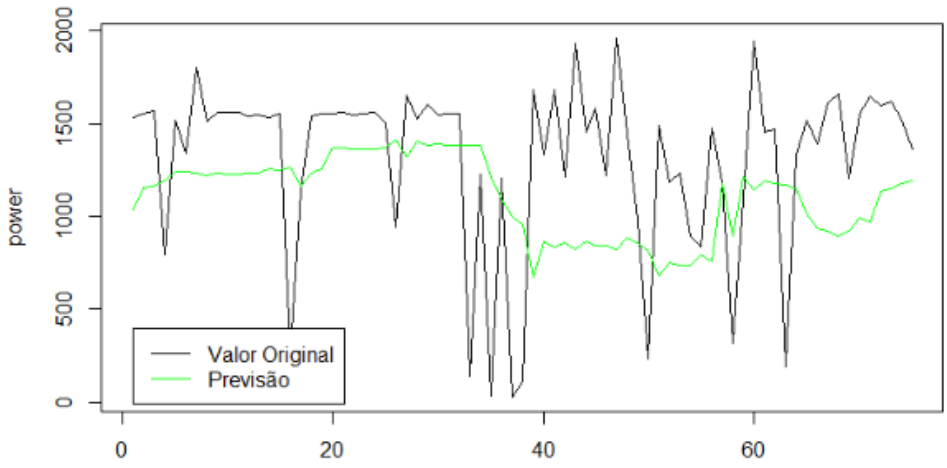

Figura 4.30: Previsão com janela de 21 modelo - AR 1.

Algo parecido acontece com os valores de erros da previsão no modelo MA, assim como no modelo AR a melhor previsão é a MA 1, isso ajuda na hipótese de que a modelagem

usando esses dois modelos não é suficiente para prever os valores de energia. Como mostrado na Figura 4.31

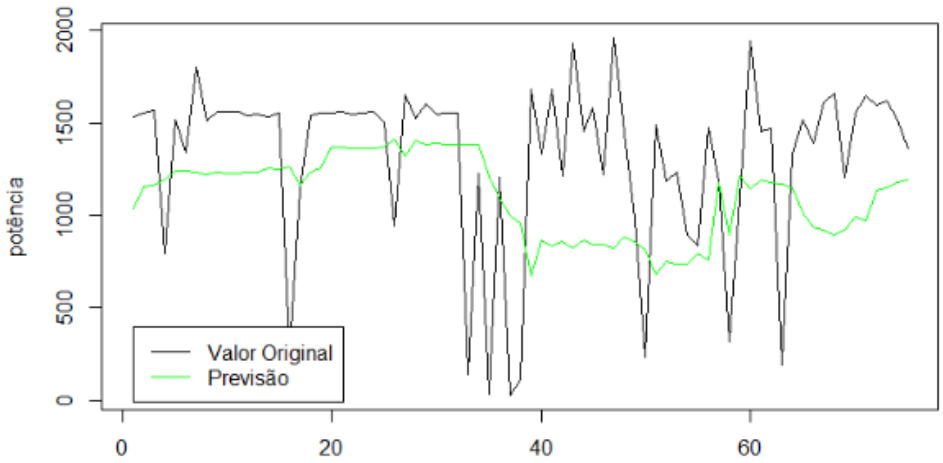

Figura 4.31: Previsão com janela de 21 modelo - MA 1.

Era esperado que os modelos mais complexos ARMA e ARIMA seriam mais capazes de prever o comportamento da série, de certa forma o modelo ARMA(2,0,1) é o de maior sucesso quando se leva em consideração o MAD. O MAD do ARMA(1,0,1) é maior que ARMA(1,0,3), essa diferença é maior ainda quando analisamos o MSE entre esses modelos, o modelo ARMA(1,0,2) tem um MSE de 1.318.600. 0 que é muito mais alto que seus vizinhos. O desequilíbrio pode ser justificado devido à janela proposta e pela variância entre os valores dos intervalos de transmissão de energia.

Apesar dos valores de previsão terem melhorado com relação ao MAD sendo mostrado o ARIMA(3,1,1), o melhor modelo com relação ao MSE entre todos ainda é o AR(1). Levando em conta a variedade de modelos junto com a variação entre os valores de MSE, acrescentando somente um valor de lag sendo p ou q pode indicar que a série atual não representa bem o conjunto dos dados ou que a complexidade da série não pode ser prevista por modelos tão simples, e por esse motivo alguma mudança deve ser feita de modo a facilitar a modelagem.

4.3.4 Modelo sem Beacon

A mudança proposta está na retirada dos beacons nos valores de transmissão da série. Os beacons mandam dados de maneira periódica durante toda a extensão da transmissão de dados. Removendo esses beacons a série de tempo teria valores com uma variância menor, além da possibilidade de remoção de um elemento repetido da série que poderia estar influenciando na modelagem.

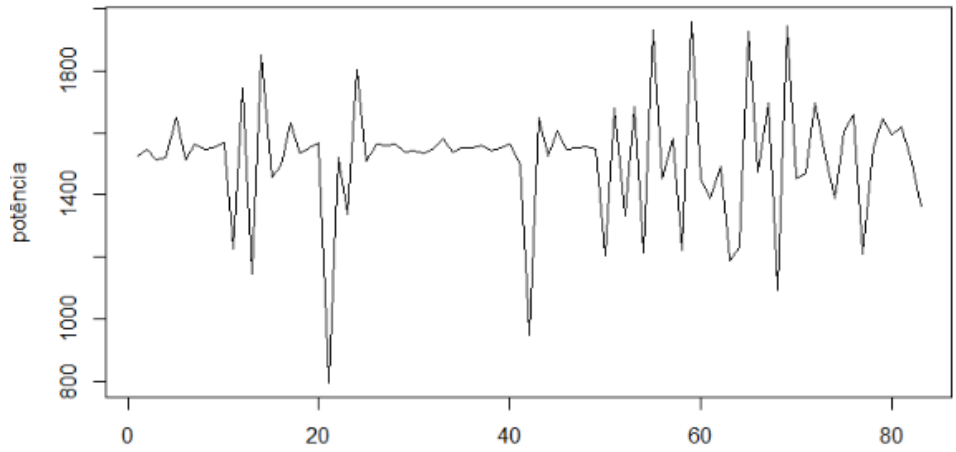

Figura 4.32: Amostra de energia sem Beacons.

A nova série é mostrada na Figura 4.32. Num total de 83 amostras que vão de um intervalo de transmissão de 795 até o maior valor de 1957. Escolhendo os primeiros 60 valores da série para um teste ACF podemos verificar uma correlação nos lags 1 e no 3 assim como no teste PACF que mostra uma correlação maior de 0,2 até o lag 3. O modelo ARIMA para essa série seria um ARIMA(3,1,1) ou um ARIMA(3,1,3).

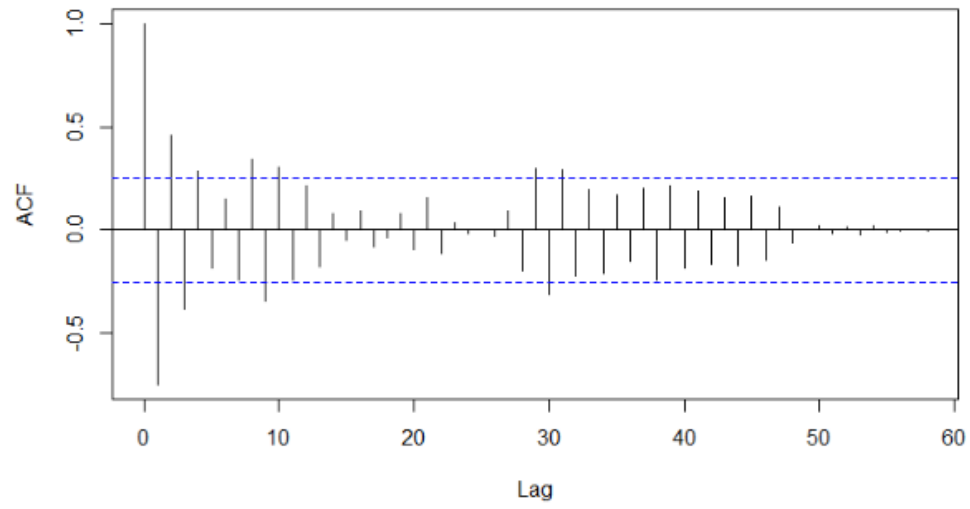

Figura 4.33: ACF dos dados sem Beacons das 60 primeiras amostras.

Foram feitas avaliações usando valores p e q entre 0 e 5 e dentro dos modelos montados o ARIMA(3,1,3) é o que apresenta o menor erro. Neste caso o modelo ARIMA(3,1,3) pode ser melhor devido à correlação evidenciados pelo teste PACF ou devido à perda de referência na parte MA do modelo de lag 1. Como pode-se ver na Figura 4.35 compa-

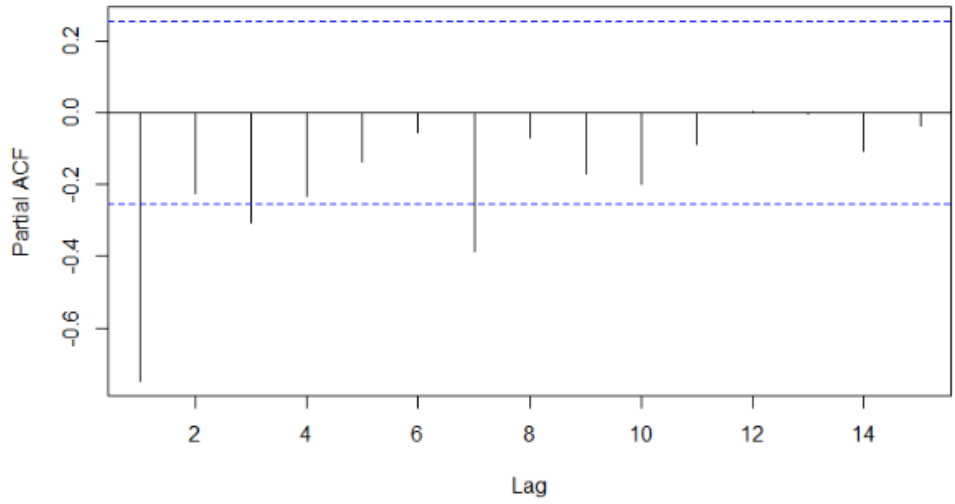

Figura 4.34: PACF dos dados sem Beacons das 60 primeiras amostras.

rada com a Figura 4.36, a diferença mais notável é a constância que a série prevista no ARIMA(3,1,1) passa a ter após aproximadamente 20 previsões.

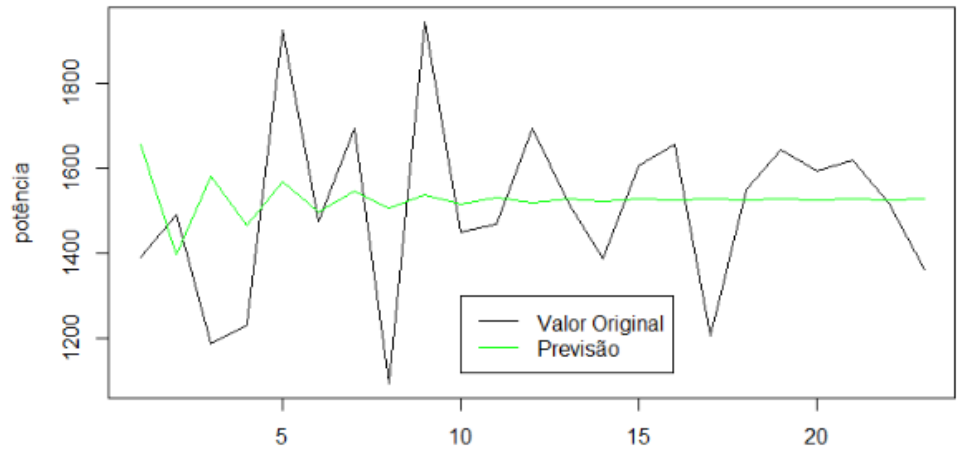

Figura 4.35: Modelos ARIMA(3,1,1) da data dos primeiros 60 valores.

Podemos concluir que para então que para melhorar os valores de previsão se faz necessário uma nova avaliação da série antes da perda de referência, que pode acarretar na série se tornando uma reta constante ou até uma função diferente como acontece na Figura 4.37 que é uma previsão utilizando os 20 primeiros valores de energia com um modelo ARIMA(9,1,2) onde após alguns valores a previsão se torna uma função periódica.

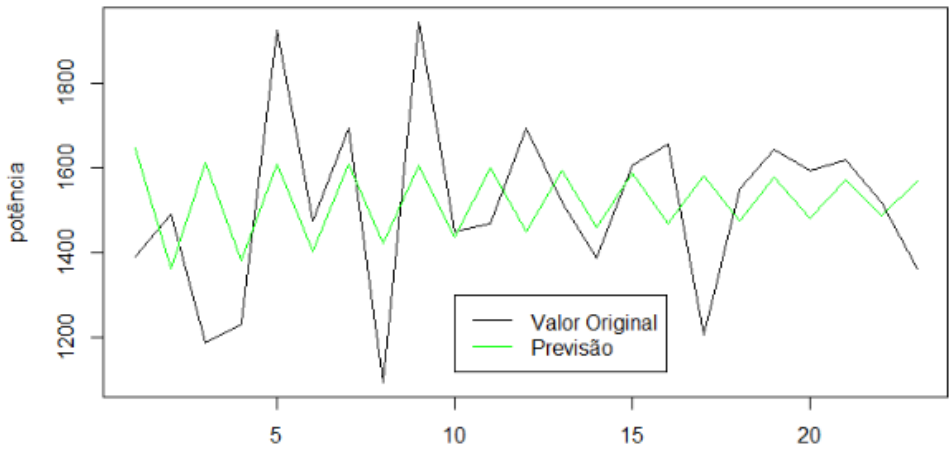

Figura 4.36: Modelos ARIMA(3,1,3) da data dos primeiros 60 valores.

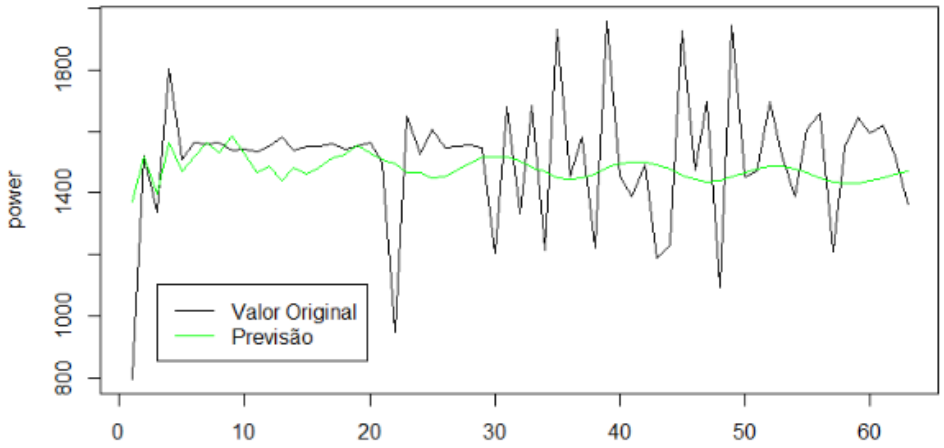

Figura 4.37: Previsão ARIMA(9,1,2) sem frequência.

4.3.5 Sem Beacon e janela de tamanho igual a 20

Outra mudança está na mudança da janela proposta anteriormente. A série com valores alterados não possui correlação até o lag 21 como a anterior, por esse motivo a solução foi escolher um número médio de previsões que melhor se aproximassem dos valores reais, além de alterar o método para gerar o modelo ARIMA.

Por comodidade e para manter a base de comparação com os exemplos anteriores, a janela usada foi de 20 valores da série, ou seja, a cada janela de 20 é feito uma previsão dos próximos 20.

Uma mudança nos resultados apresentadas na Tabela 4.14 comparando com valores anteriores é a grande diminuição dos valores MSE. Esta redução talvez possa significar uma melhora na representação da série, porém também pode ser visto que a taxa de

Tabela 4.14: Resultados MA sem beacon e janela de 20

	MSE	MAD	MPE	TC	DM
MA(1)	42165.26	137.42	10.23	36.50	198.06
MA(2)	43713.42	138.60	10.38	36.50	198.06
MA(3)	44934.53	146.98	10.87	34.92	216.31
MA(4)	45586.95	148.54	10.98	36.50	209.87

desperdício e a taxa de colisão ainda apresentam um valor muito maior do que seria aceitável em uma DSA. Esse padrão se repete para outros modelos ARIMA como pode ser visto nas Tabelas 4.15, 4.16 e 4.17.

Tabela 4.15: Resultados AR sem beacon e janela de 20

	MSE	MAD	MPE	TC	DM
AR(1)	23397.46	138.53	10.37	36.50	204.49
AR(2)	43446.15	139.01	10.40	38.09	198.04
AR(3)	43803.26	141.25	10.56	38.09	199.64
AR(4)	44176.45	142.91	10.65	38.09	200.25

Tabela 4.16: Resultados ARMA sem beacon e janela de 20

	MSE	MAD	MPE	TC	DM
ARMA(1,0,1)	43138.78	139.74	10.47	36.50	208.25
ARMA(1,0,2)	44684.59	141.80	10.64	36.50	210.49
ARMA(1,0,3)	45774.34	148.53	10.98	36.57	209.52
ARMA(1,0,4)	45926.62	149.08	11.01	36.50	210.63
ARMA(2,0,1)	43834.16	141.34	10.58	36.50	208.85
ARMA(2,0,2)	45292.74	144.82	10.83	38.10	204.30
ARMA(2,0,3)	44298.22	145.58	10.67	36.50	198.91
ARMA(2,0,4)	47892.86	152.69	11.25	36.50	213.26
ARMA(3,0,1)	42668.61	143.28	10.62	34.92	207.65
ARMA(3,0,2)	46532.51	146.43	10.93	36.50	215.38
ARMA(3,0,3)	46940.17	151.14	11.17	36.50	210.92
ARMA(3,0,4)	47759.33	152.61	11.28	36.50	213.27
ARMA(4,0,1)	42337.41	139.47	10.37	38.09	190.05
ARMA(4,0,2)	44086.92	145.52	10.79	36.50	205.87
ARMA(4,0,3)	48116.88	151.89	11.22	34.92	220.61
ARMA(4,0,4)	46182.98	149.82	11.10	36.50	208.16

Tabela 4.17: Resultados ARIMA sem beacon e janela de 20

	MSE	MAD	MPE	TC	DM
ARIMA(1,1,1)	43440.86	138.69	10.38	36.50	204.65
ARIMA(1,1,2)	45681.40	142.98	10.71	26.50	215.46
ARIMA(1,1,3)	47748.32	146.77	10.97	36.50	211.09
ARIMA(1,1,4)	47437.23	148.21	11.01	36.50	209.90
ARIMA(2,1,1)	45115.42	142.84	10.67	38.09	198.62
ARIMA(2,1,2)	46242.46	145.60	10.91	36.50	212.71
ARIMA(2,1,3)	47027.19	145.16	10.91	36.50	214.33
ARIMA(2,1,4)	55844.85	162.79	12.14	38.09	220.93
ARIMA(3,1,1)	45543.79	145.61	10.86	38.09	200.30
ARIMA(3,1,2)	47129.19	146.78	10.99	38.09	201.15
ARIMA(3,1,3)	43444.79	139.74	11.52	39.68	198.86
ARIMA(3,1,4)	51521.51	153.98	11.52	39.68	210.68
ARIMA(4,1,1)	46098.30	147.07	10.95	38.09	200.97
ARIMA(4,1,2)	48235.14	151.33	11.21	34.92	219.23
ARIMA(4,1,3)	47398.82	148.61	11.08	39.68	192.16
ARIMA(4,1,4)	52955.28	158.48	11.83	38.09	206.03

4.3.6 Perfect fit

De modo a otimizar os resultados obtidos até agora, as previsões dessa seção é a junção dos melhores modelos da seção anterior. As tabelas apresentadas até agora foram a média de todas as janelas de 20 valores, algumas delas não são as melhores para representar a série como um todo, mas é boa para representar uma única janela. Por esse motivo a melhor entre todas as previsões se originou de um modelo que otimizado para uma janela contendo poucos valores da série, talvez pelo fato da busca de oportunidades ser algo muito dinâmico e manter pouco relação com valores muito distantes.

A Figura 4.38 apresenta o conjunto de previsões feitas usando e janela de 20 e prevendo 10 valores. Analisando os resultados percebemos uma grande diferença entre o valores de erro das séries. Isso mostra a volatilidade da série, apesar dos vários testes para verificar correlação entre os números, nenhuma das previsões consegue diminuir os valores de erros apresentados em seções anteriores.

A Tabela 4.18 deixa claro a dificuldade do ARIMA para prever valores no contexto de redes, mesmo diminuindo o valor médio de desvio em alguns dos modelos como acontece nas previsões de 1-20 4.39, onde a série tem valores menos variantes, puxam muito a média de erros para baixo, enquanto previsões de 41-60 mostrada em 4.43 e de 51-70 mostrada 4.44, puxam os valores médios de erros muito para cima. Era esperado que essa variância não fosse tão grande, uma vez que uma previsão errada pode gerar uma interferência na transmissão e o intuito é de gerar oportunidades minimizando colisões.

Uma característica que todos os modelos têm em comum é a taxa de colisão que

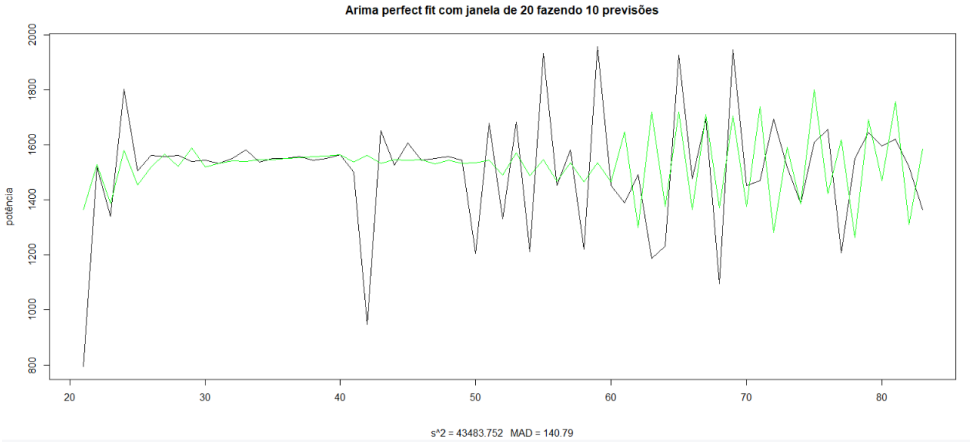

Figura 4.38: Perfect fit previsões com arimas variados.

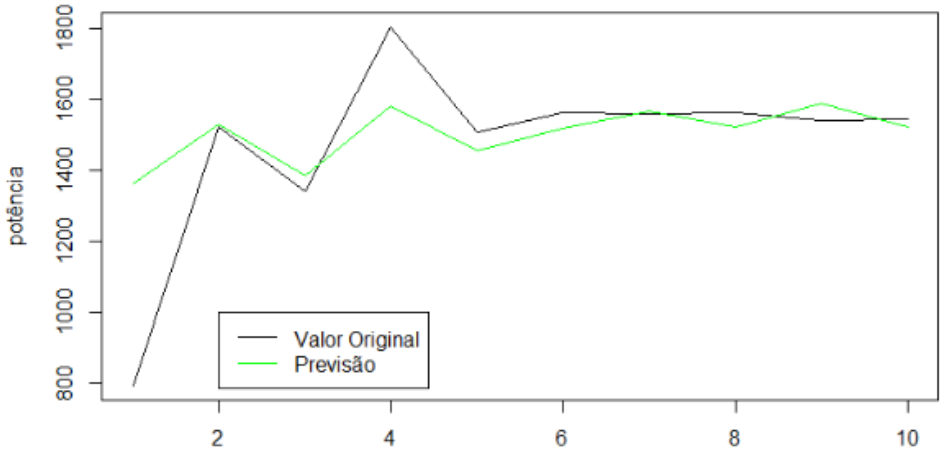

Figura 4.39: Modelos ARIMA(9,1,2) da janela 1 até 20.

Tabela 4.18: Resultados de porções diferentes das previsões

	MSE	MAD	MPE	TC	DM
1-20	38435.28	106.54	10.21	50	204.49
10-30	227.11	9.28	0.59	50	6.13
20-40	50720.06	122.91	11.01	50	200.42
30-50	52098.47	181.08	11.59	50	141.60
40-60	60167.19	204.77	14.96	60	244.88
50-70	55615.01	201.87	13.51	53.84	192.86

permanece alta mesmo com alterações feitas até agora. Nesse exemplo elas variam de 40% até 60% de erro, o que também é inaceitável no contexto de redes, onde o esperado seria de menos 2%. Obviamente os resultados apresentados caso aplicados em um protocolo,

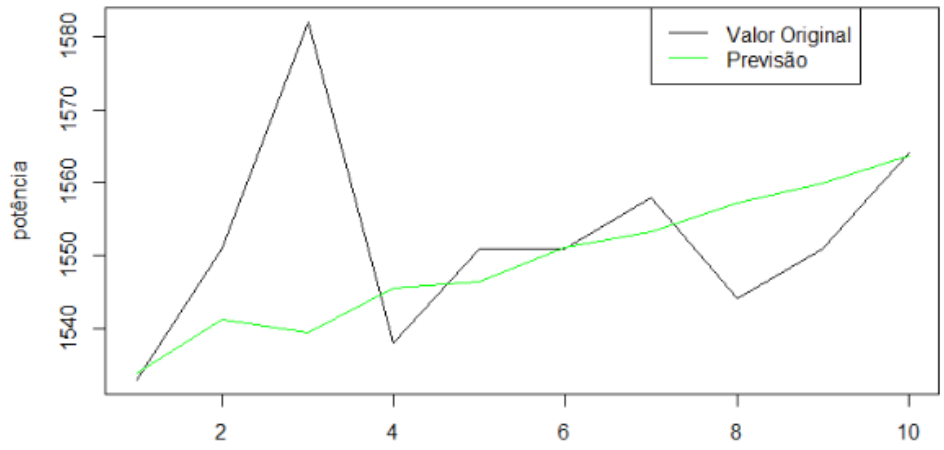

Figura 4.40: Modelos ARIMA(2,1,2) da janela 11 até 30.

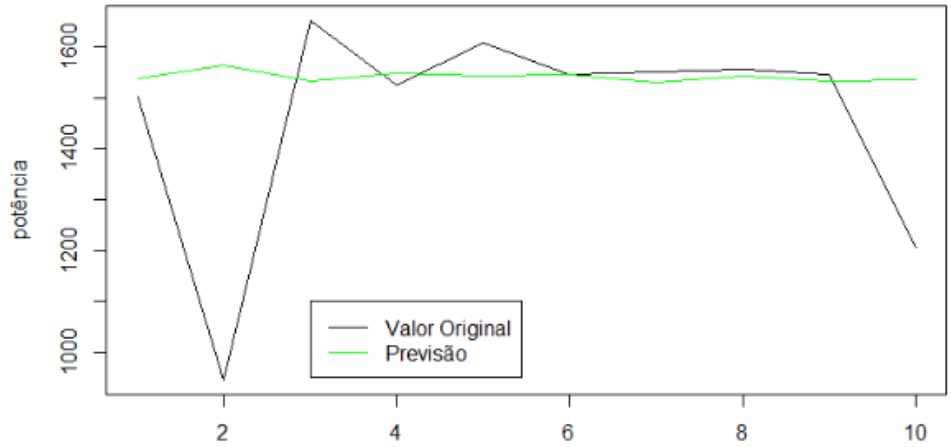

Figura 4.41: Modelos ARIMA(4,1,2) da janela 21 até 40.

diminuiriam uma vez que fossem aplicadas regras simples de transmissão, como valor mínimo para se caracterizar uma oportunidade para aquela rede específica. Mesmo assim as taxa encontradas são maiores que a maioria dos métodos disponíveis hoje.

As previsões podem apresentar uma melhora quando comparadas aos modelos antigos mas o que é notável, principalmente nas previsões de 30-50 da Figura 4.42, é que algumas tendências da série são captadas pelo ARIMA e pelo ARMA, mas que apesar dos acertos em alguns casos, o método também erra bastante, muito devido a correlação inversa de algum dos lags p ou q. Essa característica observada em várias previsões usando o modelo o torna um candidato fraco em previsões complexas, e é por isso que em muitos trabalhos o ARIMA é utilizado como coadjuvante para acesso oportunístico e para outros tipos de previsões.

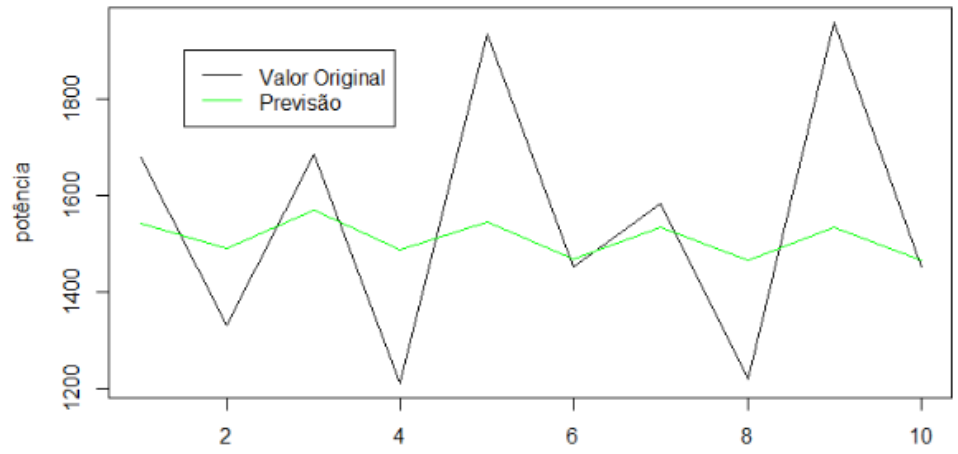

Figura 4.42: Modelos ARIMA(4,0,2) da janela 31 até 50.

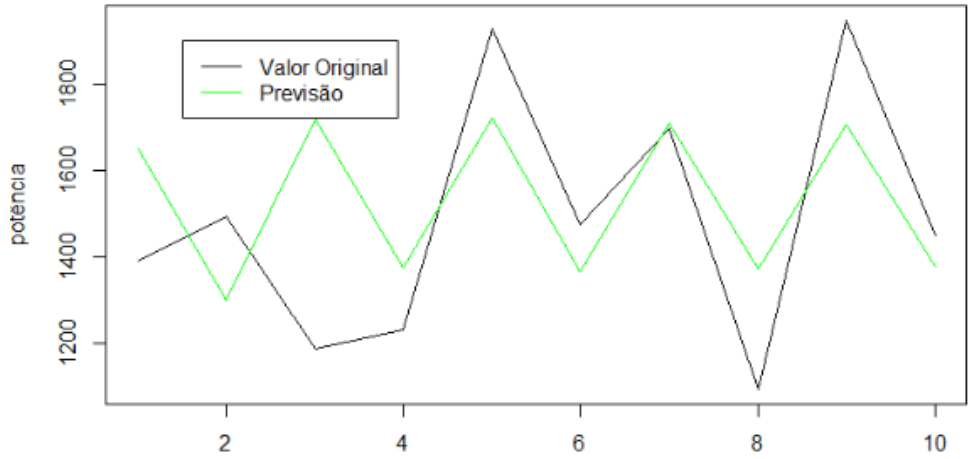

Figura 4.43: Modelos ARIMA(3,0,2) da janela 41 até 60.

Tabela 4.19: Resultado Perfect Fit

	MSE	MAD	MPE	TC	DM
Perfect$_fit$	43484.75	140.79	10.46	50.79	192.86

4.3.7 Últimos 20

Para efeito de comparação e na tentativa de melhorar as previsões, foram feitas previsões em cima dos últimos 20 valores da série.

Pensando na qualidade do modelo, o tamanho da janela foi alterado, de modo a aumentar o número de modelos. A suposição foi de que com um número grande de modelos vai melhorar ainda mais os resultados. Usando quatro tamanhos de janela diferentes teremos a previsão usando somente um modelo. Usando a janela de dez, teremos dois modelos

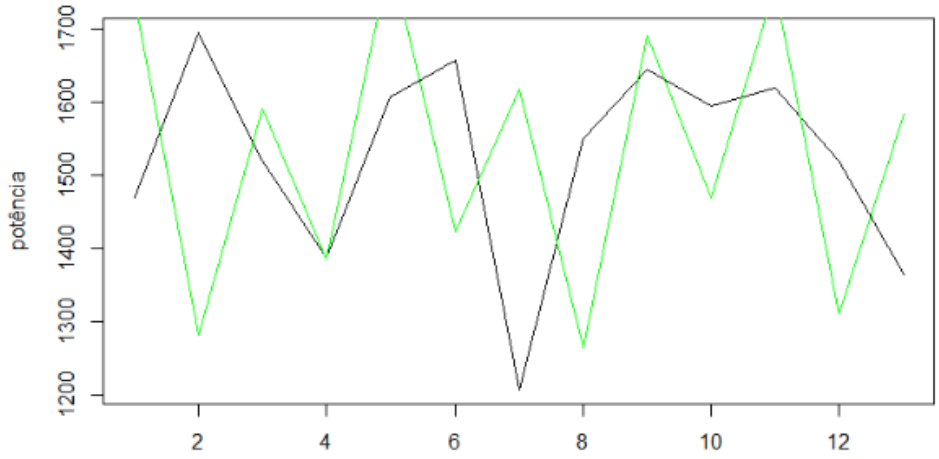

Figura 4.44: Modelos ARIMA(4,1,5) da janela 51 até 70.

ARIMAs para prever 10 valores cada, temos a janela de cinco, fazendo cinco previsões, usando quatro modelos cada e finalmente um modelo para cada previsão, totalizando 20 modelos. A Tabela 4.20 mostramos somente modelos ARMA, pois os resultados dessa tabela se repetem nas outras tabelas AR, MA e ARIMA.

Tabela 4.20: Resultados ARMA dos ultimos 20, janela com 20 previsões

	MSE	MAD	MPE	TC	DM
ARMA(1,0,1)	54514.18	179.14	12.09	40	200.94
ARMA(1,0,2)	53685.55	177.91	12.02	40	200.04
ARMA(1,0,3)	58408.15	184.63	12.55	40	222.35
ARMA(1,0,4)	58655.76	185.06	12.58	40	222.88
ARMA(2,0,1)	54240.60	178.74	12.07	40	200.63
ARMA(2,0,2)	53196.99	178.30	12.16	40	213.13
ARMA(2,0,3)	57226.31	184.26	12.53	40	221.31
ARMA(2,0,4)	59166.94	186.54	12.67	40	224.34
ARMA(3,0,1)	54483.21	180.39	12.29	40	215.82
ARMA(3,0,2)	54605.76	180.62	12.30	40	216.12
ARMA(3,0,3)	57534.28	182.86	12.42	40	219.72
ARMA(3,0,4)	57412.50	177.84	12.15	40	213.90
ARMA(4,0,1)	54561.54	180.53	12.30	40	216.02
ARMA(4,0,2)	52453.39	177.33	12.08	40	211.93
ARMA(4,0,3)	61150.91	185.28	12.62	40	222.83
ARMA(4,0,4)	62425.90	189.46	12.89	40	227.77

Ao se alterar a janela de 20 para 10 (apresentada na Tabela 4.21) existe uma piora nos valores de MSE e MAD, apesar disso o número de colisões quase não se altera. Isso pode ser explicado devido a falta de lags para previsões, e esse fenômeno não se repete para as

Tabela 4.21: resultados ARMA dos últimos 20, janela com 20 previsões

	MSE	MAD	MPE	TC	DM
ARMA(1,0,1)	54600,34	178,76	12,08	40	203,86
ARMA(1,0,2)	55978,07	180,42	12,24	40	204,51
ARMA(1,0,3)	58608,56	12,59	12,59	40	215,03
ARMA(1,0,4)	59285,43	188,07	12,73	40	217,82
ARMA(2,0,1)	56427,79	181,90	12,34	40	207,07
ARMA(2,0,2)	55296,00	181,46	12,37	40	208,61
ARMA(2,0,3)	57580,55	186,96	12,66	40	215,99
ARMA(2,0,4)	61668,40	192,16	13,05	40	223,38
ARMA(3,0,1)	54700,77	182,51	12,38	40	209,62
ARMA(3,0,2)	56953,69	184,44	12,57	40	213,48
ARMA(3,0,3)	57814,98	185,49	12,55	40	214,04
ARMA(3,0,4)	59366,36	183,47	12,53	40	212,68
ARMA(4,0,1)	54784,32	182,66	12,39	40	209,84
ARMA(4,0,2)	54635,56	182,68	12,44	40	210,76
ARMA(4,0,3)	67817,35	202,73	13,78	45	209,84
ARMA(4,0,4)	64237,21	195,90	13,31	40	227,31

próximas tabelas. Talvez com um número maior de p e q as previsões com a janela de 20 seriam melhores na média que as previsões com a janela de 10.

Como o tamanho da janela está menor os valores previstos apresentam resultados melhores de MSE e MAD, como mostrado na Tabela 4.22. Apesar disso os valores de taxa de colisão permanecem em 40, o que pode ser um problema no contexto de redes.

Quando escolhemos os modelos diferentes é possível delinear ainda mais os resultados. Algumas outras combinações de ARMA e ARIMA foram feitas na Figura 4.45. Nessa imagem são mostrados os resultados da combinação de modelos ARIMA(3,0,14) ARIMA(3,0,4) ARIMA(1,1,1) e ARIMA(4,1,3), que contém valores MSE menores que as outras tabelas apresentadas nessa seção até agora.

Tabela 4.22: Tabela ARMA resultados dos últimos 20, janela com 20 previsões

	MSE	MAD	MPE	DM	TC
ARMA(1,0,1)	53996,96	189,81	12,89	199,00	45
ARMA(1,0,2)	52712,56	184,21	12,59	219,26	40
ARMA(1,0,3)	54746,61	184,67	12,64	208,10	45
ARMA(1,0,4)	55540,85	187,42	12,83	238,17	40
ARMA(2,0,1)	53683,00	187,4	12,79	222,30	40
ARMA(2,0,2)	52217,76	183,62	12,63	205,41	45
ARMA(2,0,3)	52508,84	182,22	12,50	184,70	50
ARMA(2,0,4)	59415,34	196,41	13,45	221,55	45
ARMA(3,0,1)	49692,39	178,99	12,31	201,37	45
ARMA(3,0,2)	53892,85	186,22	12,80	210,15	45
ARMA(3,0,3)	54457,58	182,06	12,45	184,21	50
ARMA(3,0,4)	48851,21	162,52	11,34	206,65	40
ARMA(4,0,1)	50860,45	182,2	12,51	204,66	45
ARMA(4,0,2)	52203,83	186,7	12,84	210,74	45
ARMA(4,0,3)	59321,6	186,67	12,9	209,00	45
ARMA(4,0,4)	62010,21	197,11	13,53	204,39	50

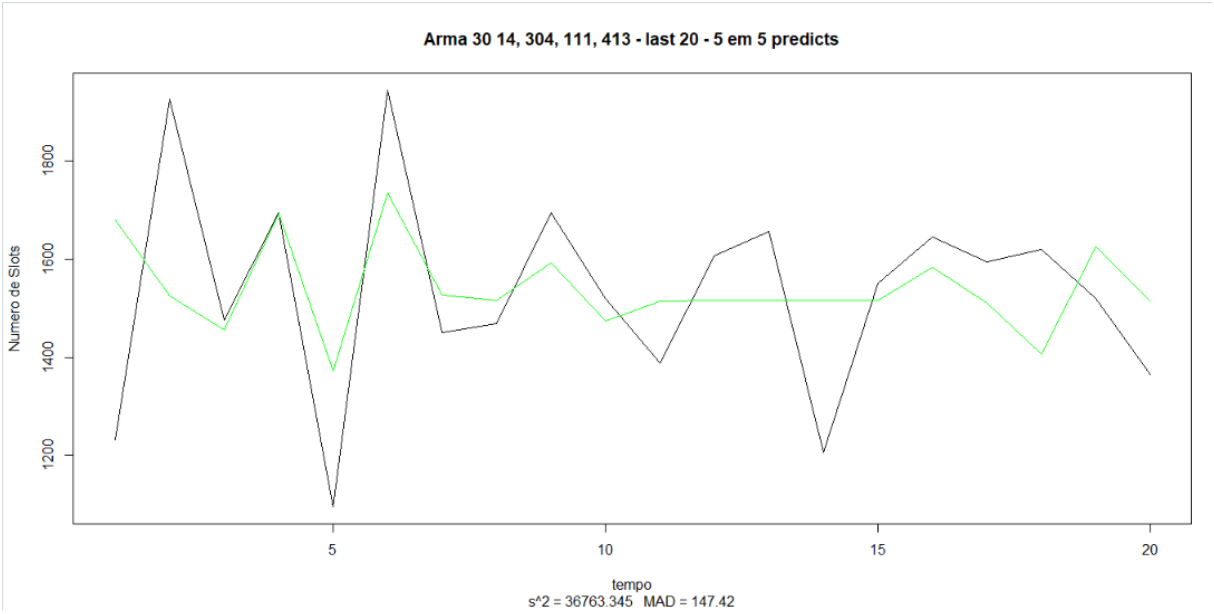

Figura 4.45: Gráfico com Armas variados com janela de 5.

Finalmente os resultados com os últimos 20 apresentam os menores MSE na Tabela 4.23, nesse caso foram 20 modelos ARMA com somente uma previsão, e nesse caso, o valor da taxa de colisão caiu para 35. Um dos problemas dessa abordagem é a dificuldade em detectar os valores p e q para cada previsão e a dificuldade na implementação desse modelo num ambiente real.

Tabela 4.23: Resultados ARMA dos últimos 20, janela com 20 previsões

	MSE	MAD	MPE	DM	TC
ARMA(1,0,1)	42611,59	180,31	12,23	222,96	35
ARMA(1,0,2)	37776,31	164,54	11,21	197,64	35
ARMA(1,0,3)	41751,78	164,05	11,30	206,30	35
ARMA(1,0,4)	44385,97	173,00	11,86	208,03	35
ARMA(2,0,1)	37659,76	169,04	11,38	179,53	35
ARMA(2,0,2)	34657,32	149,30	10,24	174,84	35
ARMA(2,0,3)	39640,14	162,15	11,16	198,29	35
ARMA(2,0,4)	39773,68	160,42	11,06	196,71	35
ARMA(3,0,1)	37478,97	161,68	11,09	192,72	35
ARMA(3,0,2)	38947,51	160,52	11,01	184,88	35
ARMA(3,0,3)	39579,02	155,30	10,75	163,58	35
ARMA(3,0,4)	41437,68	154,35	10,73	186,17	35
ARMA(4,0,1)	38566,65	164,54	11,28	195,87	35
ARMA(4,0,2)	36322,00	155,85	10,69	184,72	35
ARMA(4,0,3)	46070,21	169,77	11,73	205,71	35
ARMA(4,0,4)	41466,10	156,48	10,85	191,93	35

Com relação aos dados apresentados pode-se perceber que ainda existe uma grande variância com relação a modelagem. Observando um exemplo dos melhores resultados do modelo com uma janela de 5 é possível notar que existe uma diferença na qualidade das previsões dependendo da porção que se está avaliando. Isso se dá devido da complexidade dos dados, para algumas porções o melhor modelo ARIMA pode ser diferente de alguma outra porção. O melhor resultado entre todos as análises do terceiro conjunto de dados é apresentado na Figura 4.46

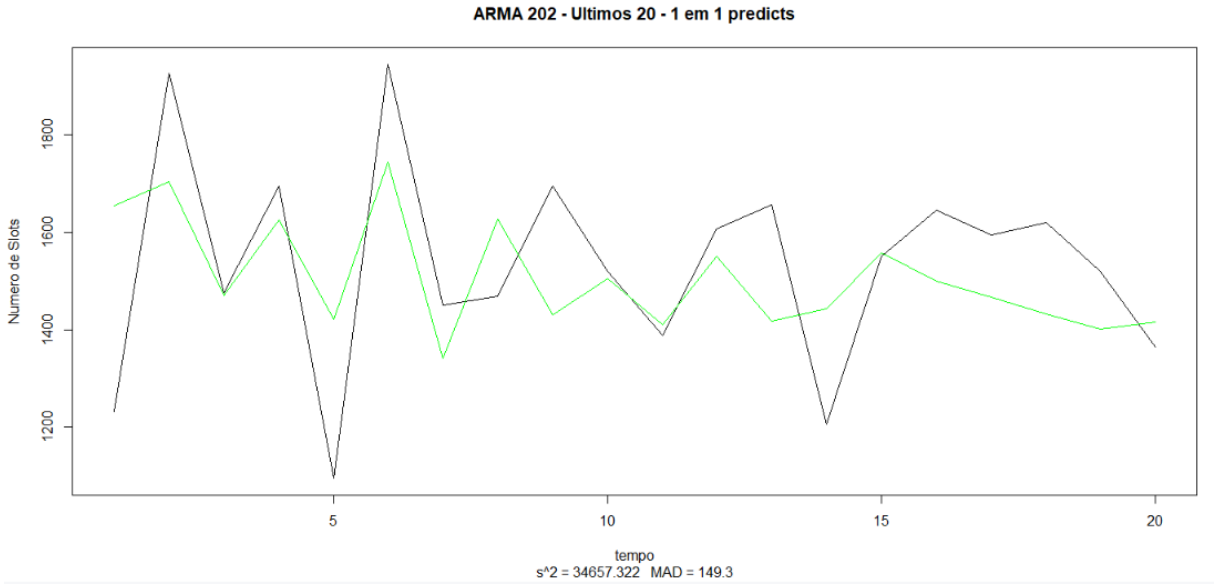

Figura 4.46: Gráfico ARMA(202) menor MSE com janela de previsão unitária.

Capítulo 5

Conclusão

No contexto da comunicação sem fios, existe a necessidade de aumentar a utilização do canal de comunicação. Para isso é possível, usar do acesso oportunístico, de modo a permitir que usuários secundários. Para isso é necessário caracterizar o comportamento de uso do usuário primário, gerando oportunidades de acesso ao canal. Nesse trabalho foi mostrado uma variedade de resultados referentes a previsões que utilizam modelos lineares. O objetivo por trás desse estudo foi de descobrir se esses modelos lineares seriam capazes de serem usados para prever o comportamento do usuário primário.

5.1 Resultados finais

Com uma variedade de cenários apresentados nas seções anteriores foi possível observar algumas das capacidades do modelo ARIMA e testar sua aplicação em cenários bastante distintos.

Os dois primeiros cenários apresentados(*Airpassengers* e Netflix) possuem uma variança menor que o cenário montado usando as capturas de rede. Essa característica pode justificar a diferença nos resultados e a dificuldade do ARIMA em fazer previsões nesse modelo. Outra característica que é notada nos dois primeiros modelos e não é notada no modelo final é a sazonalidade, que impacta muito nos resultados.

5.2 Trabalhos futuros

Em resumo, os dados reunidos para esse estudo são complexos demais para um modelo tão simples como o ARIMA, talvez o uso de SARIMA [23] para modelagem pudesse ser feita, desse modo componentes sazonais pudessem ser incluídas e assim englobando correlações mais complexas.

A diferença de modelagem alterando entre modelos eficientes a curto prazo e a constante refatoração, comparado a modelos menos eficientes e com menor capacidade de previsão também foi estudada.

A construção da série de tempo influencia os resultados do modelo ARIMA. Um estudo sobre a construção da série poderia ser feita, onde o foco série na periodicidade da aquisição dos dados e não na contagem de *slots idle* na transmissão Wifi. Também poderia ser pensado alguma outra aplicação para o uso do ARIMA, um sistema que aceitaria um erro maior ou que fosse mais fácil adquirir dados.

Referências

[1] ANATEL: *Quadro de atribuição de faixas de frequências no brasil*. https://espectro.org.br/pt-br/content/plano-de-atribuição-de-faixas-de-frequência-no-brasil-2018, 2018. vii, 2

[2] Moraes Modesto, Felipe de: *Um protocolo de acesso ao meio baseado na análise de disponibilidade do espectro*. Departamento de Ciência da Computação, 2011. vii, 7

[3] James F. Kurose, Keith W. Ross: *Redes de computadores e a internet uma abordagem top-down*, volume 6a edição. Pearson Education, 2003, ISBN 978-85-430-1443-2. vii, 11, 12, 14

[4] Zhao, Qing e Brian M Sadler: *A survey of dynamic spectrum access*. Signal Processing Magazine, páginas 79–89, 2007. vii, 15, 16

[5] RakanNimer: *Air passengers*, 2016. vii, 22, 23

[6] Shao, Junyan: *Web traffic time series prediction using arima lstm*, 2019. `https://medium.com/@jyshao53/web-traffic-time-series-prediction-using-arima-lstm-7ef3911845ae`, acesso em 2020-15-12. vii, 23, 34, 35

[7] Matthew L. Smith, Randy Spence, Ahmed T. Rashid: *Mobile phones and expanding human capabilities*. Mobile Telephony Special Issue, Volume 7(Number 03):77–88, 2011. 1

[8] Cisco: *Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022*. Cisco public, February, 2019. 1

[9] R. K. Agrawal, Ratnadip Adhikari e: *An Introductory Study on Time series Modeling and Forecasting*, volume v. 1. LAP Lambert Academic Publishings, Janeiro 2013. 5

[10] Jonathan D. Cryer, Kung Sik Chan: *Time Series Analysis With Applications in R*, volume second edition. Springer, 2008. 6, 8

[11] David, F. N. *Biometrika*, 39(1/2):213–214, 1952, ISSN 00063444. `http://www.jstor.org/stable/2332487`. 8

[12] R. A. Davis, P. J. Brockwell e: *Time series: Theory and methods*. 2(2):220–225, 373–375, 1991. 10

[13] Claudia Cormio, Kaushik R. Chowdhury: *A survey on mac protocols for cognitive radio networks*. Ad Hoc Networks, 7, February 2009. 14

[14] Janez Sterle, Mojca Volk, Urban Sedlar Janez Bester e Andrej Kos: *Application-based ngn qoe controller*. IEEE Communications Magazine, Volume 49(Number 1):92 – 101, 2011. 16

[15] ucek, Tevfik Y e H useyin Arslan: *A survey of spectrum sensing algorithms for cognitive radio applications*. IEEE COMMUNICATIONS SURVEYS TUTORIALS, VOL. 11(NO. 1):116 – 130, FIRST QUARTER 2009. 17

[16] Arslan(Ed.)., Hüseyin: *Cognitive Radio, Software Defined Radio, and Adaptive Wireless Systems.* Springer, 2007. 18

[17] Hongjun Wang, Kaiying Wang, Hui Zhao e Youjun Yue: *Prediction of user behavior in smart home based on improved arima model*. International Conference on Mechatronics and Automation, Changchun, China, 2018, August 5 - 8. 19

[18] Rania T. Fleifel, Samy S. Soliman, Walaa Hamouda Ashraf Badawi: *Lte primary user modeling using a hybrid arima/narx neural network model in cr*. 2017 IEEE Wireless Communications and Networking Conference (WCNC), 19-22 March 2017. 20

[19] Zhang, G. Peter: *Time series forecasting using a hybrid arima and neural network model*. Neurocomputing 50 (2003) 159 – 175, 16 July 1999. 20

[20] Cao, L.J. e Francis E.H. Tay: *Support vector machine with adaptive parameters in financial time series forecasting*. IEEE Transaction on Neural Networks, Vol. 14. 21

[21] Google: *Web traffic time series forecasting*. `https://www.kaggle.com/c/web-traffic-time-series-forecasting`, acesso em 2020-01-04. 23, 34

[22] Hawkins, Douglas M.: *The problem of overfitting*, volume 44 (1): 1–12. Journal of Chemical Information and Modeling, 2004. 41

[23] Rob J Hyndman, George Athanasopoulos: *Forecasting: principles and practice*, volume 1a edição. OTexts, 2018, ISBN 978-0-9875071-1-2. 67